

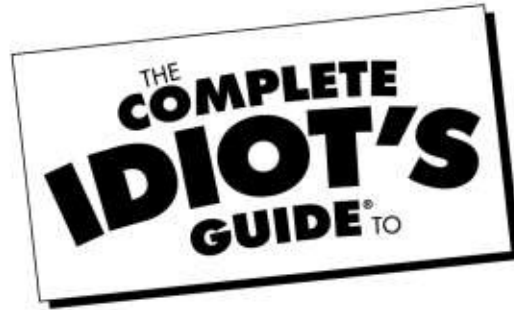
THE
**COMPLETE
IDIOT'S**
GUIDE[®] TO

Statistics

SECOND EDITION

Simple solutions to the biggest statistical
problem—getting through the class

Robert A. Donnelly, Jr., Ph.D.



Statistics

Second Edition

by Robert A. Donnelly, Jr., Ph.D.



A member of Penguin Group (USA) Inc.

To my wife, Debbie, who supported and encouraged me every step of the way. I could not have done this without you, babe.

ALPHA BOOKS

Published by the Penguin Group

Penguin Group (USA) Inc., 375 Hudson Street, New York, New York
10014, U.S.A.

Penguin Group (Canada), 10 Alcorn Avenue, Toronto, Ontario, Canada
M4V 3B2 (a division of Pearson Penguin

Canada Inc.)

Penguin Books Ltd, 80 Strand, London WC2R ORL, England

Penguin Ireland, 25 St Stephen's Green, Dublin 2, Ireland (a division of
Penguin Books Ltd)

Penguin Group (Australia), 250 Camberwell Road, Camberwell, Victoria
3124, Australia (a division of Pearson

Australia Group Pty Ltd)

Penguin Books India Pvt Ltd, 11 Community Centre, Panchsheel Park,
New Delhi—110 017, India

Penguin Group (NZ), cnr Airborne and Rosedale Roads, Albany,
Auckland 1310, New Zealand (a division of

Pearson New Zealand Ltd)

Penguin Books (South Africa) (Pty) Ltd, 24 Sturdee Avenue, Rosebank,
Johannesburg 2196, South Africa

Penguin Books Ltd, Registered Offices: 80 Strand, London WC2R ORL,
England

Copyright © 2007 by Robert A. Donnelly, Jr.

All rights reserved. No part of this book shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Neither is any liability assumed for damages resulting from the use of information contained herein. For information, address Alpha Books, 800 East 96th Street, Indianapolis, IN 46240. THE COMPLETE IDIOT'S GUIDE TO and Design are registered trademarks of Penguin Group (USA) Inc. ISBN: 1-4295-1390-X Library of Congress Catalog Card Number: 2006938600

Interpretation of the printing code: The rightmost number of the first series of numbers is the year of the book's printing; the rightmost number of the second series of numbers is the number of the book's printing. For example, a printing code of 07-1 shows that the first printing occurred in 2007.

Note: This publication contains the opinions and ideas of its author. It is intended to provide helpful and informative material on the subject matter covered. It is sold with the understanding that the author and publisher are not engaged in rendering professional services in the book. If the reader requires personal assistance or advice, a competent professional should be consulted.

The author and publisher specifically disclaim any responsibility for any liability, loss, or risk, personal or otherwise, which is incurred as a consequence, directly or indirectly, of the use and application of any of the contents of this book.

Publisher: Marie Butler-Knight Editorial Director: Mike Sanders
Managing Editor: Billy Fields Acquisitions Editor: Tom Stevens
Development Editor: Michael Thai Production Editor: Kayla Dagger Copy
Editor: Nancy Wagner

Cartoonist: Chris Eltopalos Cover Designer: Bill Thomas Book Designer:
Trinn Wurst Indexer: Angle Bess Layout: Chad Dressier Proofreader:
Aaron Black

Contents at a Glance

Part I: The Basics I

1 Let's Get Started 3 Statistics plays a vital role in today's society by providing the foundation for sound decisions.

2 Data, Data Everywhere and Not a Drop to Drink 15 All statistical analysis begins with the proper selection of the source, type, and measurement scale of the data.

3 Displaying Descriptive Statistics 29

A vast array of methods display data and information effectively, such as frequency distributions, histograms, pie charts, and bar charts.

4 Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 47

Using the mean, median, or mode is an effective way to summarize many pieces of data.

5 Calculating Descriptive Statistics: Measures of Dispersion 61

The standard deviation, range, and quartiles reveal valuable information about the variability of the data.

Part 2: Probability Topics 79

6 Introduction to Probability 81

Basic probability theory, such as the intersection and union of events, provides important groundwork for statistical concepts.

7 More Probability Stuff 93 Calculate the probability of whining your tennis match given

that you had a short warm-up period.

8 Counting Principles and Probability Distributions 105 Determine your odds at winning a state lottery drawing or

your chances of drawing a five-card flush in poker.

9 The Binomial Probability Distribution 121 Calculate the probability of correctly guessing the answer of 6

out of 12 multiple-choice questions when each question has five choices.

10 The Poisson Probability Distribution 131 Determine the probability

that you will receive at least 3

spam, e-mails tomorrow given that you average 2.5 such e-mails per day.

11 The Normal Probability Distribution 145 Determine probabilities of events that follow this symmetrical, bell-shaped distribution.

Part 3: Inferential Statistics 163

12 Sampling 165 Discover how to choose between simple random, systematic, cluster, and stratified sampling for statistical analysis.

13 Sampling Distributions 177 The central limit theorem tells us that sample means follow

the normal probability distribution as long as the sample size is large enough.

14 Confidence Intervals 195 A confidence interval is a range of values used to estimate a population parameter.

15 Introduction to Hypothesis Testing 213 A hypothesis test enables us to investigate an assumption about

a population parameter using a sample.

16 Hypothesis Testing with One Sample 227 This procedure focuses on testing a statement concerning a

single population.

17 Hypothesis Testing with Two Samples 249 Use this test to see whether that new golf instructional video

will really lower your scores.

Part 4: Advanced Inferential Statistics 271

18 The Chi-Square Probability Distribution 273 This procedure enables us to test the independence of two categorical variables.

19 Analysis of Variance 289 Learn how to test the difference between more than two population means.

20 Correlation and Simple Regression 309 Determine the strength and direction of the linear relationship between an independent and dependent variable.

Contents at a Glance

Appendixes

A Solutions to "Your Turn" 333

B Statistical Tables 367

C Glossary 377

Contents

Part I: The Basics 1

1 Let's Get Started 3

Where Is This Stuff Used? 4

Who Thought of This Stuff? 5

Early Pioneers 5

More Recent Famous People 6

The Field of Statistics Today 6

Descriptive Statistics — the Minor League 7

Inferential Statistics — the Major League 8

Ethics and Statistics—It's a Dangerous World Out There 10

Your Turn 12

2 Data, Data Everywhere and Not a Drop to Drink 15

The Importance of Data 16

The Sources of Data—Where Does All This Stuff Come
From? 17

Direct Observation — I'll Be Watching You 19

Experiments — Who's in Control? 19

Sm-veys — Is That Your Final Answer? 20

Types of Data 20

Types of Measurement Scales—a Weighty Topic 21

Nominal Level of Measurement 21

Ordinal Level of Measurement 21

Interval Level of Measurement 22

Ratio Level of Measurement 22

Computers to the Rescue 23

The Role of Computers in Statistics 23

Installing the Data Analysis Add-In 24

Your Turn 26

3 Displaying Descriptive Statistics 29

Frequency Distributions 30

Constructing a Frequency Distribution 31

(A Distant) Relative Frequency Distribution 32

Cumulative Frequency Distribution 33

Graphing a Frequency Distribution — the Histogram, 34

viii The Complete Idiot's Guide to Statistics, Second Edition

Statistical Flower Power—the Stem and Leaf Display 37

Charting Your Course 39

What's Your Favorite Pie Chart? 39

Bar Charts 41

Line Charts 43

Your Turn 44

k Calculating Descriptive Statistics: Measures of Central Tendency

(Mean, Median, and Mode) k1

Measures of Central Tendency 48

Mean 48

Weighted Mean 50

Mean of Grouped Data from a Frequency Distribution 51

Median 54

Mode 55

How Does One Choose? 56

Using Excel to Calculate Central Tendency 56

Your Turn 58

5 Calculating Descriptive Statistics: Measures of Dispersion 61

Range 62

Variance 63

Using the Raw Score Method (When Grilling) 64

The Variance of a Population 65

Standard Deviation 67

Calculating the Standard Deviation of Grouped Data 67

The Empirical Rule: Working the Standard Deviation 69

Chebyshev's Theorem 71

Measures of Relative Position 73

Quartiles 13

Interquartile Range 14

Using Excel to Calculate Measures of Dispersion 75

Your Turn 76

Part 2: Probability Topics 79

6 Introduction to Probability 81

What Is Probability? 82

Contents jx

Empirical Probability 83

Subjective Probability 85

Basic Properties of Probability 86

The Intersection of Events 87

The Union of Events: A Marriage Made in Heaven 88

Your Turn 89

7 More Probability Stuff 93

Conditional Probability 94

Independent Versus Dependent Events 96

Multiplication Rule of Probabilities 97

Mutually Exclusive Events 98

Addition Rule of Probabilities 99

Summarizing Our Findings 101

Bayes' Theorem 102

Your Turn 103

8 Counting Principles and Probability Distributions 105

Counting Principles 106

The Fundamental Counting Principle 106

Permutations 707

Combinations 109

Using Excel to Calculate Permutations and Combinations III

Probability Distributions 112

Random Variables 112

Discrete Probability Distributions 113

Rules for Discrete Probability Distributions 115

The Mean of a Discrete Probability Distribution 115

The Variance and Standard Deviation of a Discrete Probability Distribution 116

Your Turn 118

9 The Binomial Probability Distribution 121

Characteristics of a Binomial Experiment 122

The Binomial Probability Distribution 123

Binomial Probability Tables 126

Using Excel to Calculate Binomial Probabilities 127

The Mean and Standard Deviation for the Binomial Distribution 129

The Complete Idiot's Guide to Statistics, Second Edition

10 The Poisson Probability Distribution 131

Characteristics of a Poisson Process 132

The Poisson Probability Distribution 133

Poisson Probability Tables 136

Using Excel to Calculate Poisson Probabilities 139

Using the Poisson Distribution as an Approximation to
the Binomial Distribution 140

Your Turn 142

11 The Normal Probability Distribution 146

Characteristics of the Normal Probability Distribution 146

Calculating Probabilities for the Normal Distribution 148

Calculating the Standard Z-Score 148

Using the Standard Normal Table 750

The Empirical Rule Revisited 755

Calculating Normal Probabilities Using Excel 756

Using the Normal Distribution as an Approximation to the Binomial
Distribution 157

Your Turn 161

Part 3: Inferential Statistics 163

12 Sampling 165

Why Sample? 166

Random Sampling 167

Simple Random Sampling 168

Systematic Sampling 770

Cluster Sampling 777

Stratified Sampling 772

Sampling Errors 173

Examples of Poor Sampling Techniques 174

Your Turn 176

13 Sampling Distributions 177

What Is a Sampling Distribution? 177

Sampling Distribution of the Mean 178

The Central Limit Theorem 182

Contents xi

Why Does the Central Limit Theorem Work? 186

Putting the Central Limit Theorem to Work 188

Sampling Distribution of the Proportion 190

Calculating the Sample Proportion 190

Calculating the Standard Error of the Proportion 192

Your Turn 193

Confidence Intervals 195

Confidence Intervals for the Mean with Large Samples 196

Estimators 196

Confidence Levels 191

Beware of the Interpretation of Confidence Interval! 199

The Effect of Changing Confidence Levels	200
The Effect of Changing Sample Size	201
Determining Sample Size for the Mean	202
Calculating a Confidence Interval When σ Is Unknown	202
Using Excel's CONFIDENCE Function	203
Confidence Intervals for the Mean with Small Samples	204
When σ Is Known	204
When σ Is Unknown	205
Confidence Intervals for the Proportion with Large Samples	208
Calculating the Confidence Interval for the Proportion	209
Determining Sample Size for the Proportion	210
Your Turn	211
IS Introduction to Hypothesis Testing	ZI3
Hypothesis Testing—the Basics	214
The Null and Alternative Hypothesis	215
Stating the Null and Alternative Hypothesis	216
Two-Tail Hypothesis Test	277
One-Tail Hypothesis Test	218
Type I and Type II Errors	219
Example of a Two-Tail Hypothesis Test	220

Using the Scale of the Original Variable	221
Using the Standardized Normal Scale	222
Example of a One-Tail Hypothesis Test	223
xii The Complete Idiot's Guide to Statistics, Second Edition	
16 Hypothesis Testing with One Sample	227
Hypothesis Testing for the Mean with Large Samples	228
When Sigma Is Known	228
When Sigma Is Unknown	229
The Role of Alpha in Hypothesis Testing	231
Introducing the p-Value	233
The p-Value for a One-Tail Test	233
The p-Value for a Two-Tail Test	234
Hypothesis Testing for the Mean with Small Samples	236
When Sigma Is Known	236
When Sigma Is Unknown	257
Using Excel's TINV Function	241
Hypothesis Testing for the Proportion with Large Samples....	242
One-Tail Hypothesis Test for the Proportion	243
Two-Tail Hypothesis Test for the Proportion	245
Your Turn	246
IT Hypothesis Testing with Two Samples	249

The Concept of Testing Two Populations 250

Sampling Distribution for the Difference in Means 250

Testing for Differences Between Means with Large

Sample Sizes 252

Testing a Difference Other Than Zero 255

Testing for Differences Between Means with Small

Sample Sizes and Unknown Sigma 256

Equal Population Standard Deviations 257

Unequal Population Standard Deviations 260

Letting Excel Do the Grunt Work 261

Testing for Differences Between Means with Dependent

Samples 263

Testing for Differences Between Proportions with

Independent Samples 265

Your Turn 269

Part 4: Advanced Inferential Statistics 271

18 The (hi-Square Probability Distribution 273

Review of Data Measurement Scales 274

The Chi-Square Goodness-of-Fit Test 274

Contents xiii

Observed Versus Expected Frequencies 276

Calculating the Chi-Square Statistic	277
Determining the Critical Chi-Square Score	277
Using Excel's CHINVFFunction	279
Characteristics of a Chi-Square Distribution	279
A Goodness-of-Fit Test with the Binomial Distribution	280
Chi-Square Test for Independence	282
Your Turn	286
19 Analysis of Variance	289
One-Way Analysis of Variance	290
Completely Randomized ANOVA	291
Partitioning the Sum of Squares	292
Determining the Calculated F-Statistic	295
Determining the Critical F-Statistic	296
Using Excel to Perform One-Way ANOVA	298
Pairwise Comparisons	299
Completely Randomized Block ANOVA	301
Partitioning the Sum of Squares	302
Determining the Calculated F-Statistic	303
To Block or Not to Block, That Is the Question	304
Your Turn	305
20 Correlation and Simple Regression	309

Independent Versus Dependent Variables 310

Correlation 311

Correlation Coefficient 312

Testing the Significance of the Correlation Coefficient 314

Using Excel to Calculate Correlation Coefficients 315

Simple Regression 316

The Least Squares Method 311

Confidence Interval for the Regression Line 321

Testing the Slope of the Regression Line 323

The Coefficient of Determination 324

Using Excel for Simple Regression 325

A Simple Regression Example -with Negative Correlation 326

Assumptions for Simple Regression 330

Simple Versus Multiple Regression 330

Your Turn 331

Appendixes

A Solutions to "Your Turn" 333

B Statistical Tables 367

(Glossary 377

Foreword

Statistics, statistics everywhere, but not a single word can we understand!

Actually, understanding statistics is a critically important skill that we all need to have in this day and age. Every day, we are inundated with data about politics, sports, business, the stock market, health issues, financial matters, and many other topics. Most of us don't pay much attention to most of the statistics we hear, but more importantly, most of us don't really understand how to make sense of the numbers, ratios, and percentages with which we are constantly barraged. In order to obtain the truth behind the numbers, we must be able to ascertain what the data is really saying to us. We need to determine whether the data is biased in a particular direction or whether the true, balanced picture is correctly represented in the numbers. That is the reason for reading this book.

Statistics, as a field, is usually not the most popular topic or course in school. In fact, many people will go to great lengths to avoid having to take a statistics course. Many people think of it as a math course or something that is very quantitative, and that scares them away. Others, who get past the math, do not have the patience to search for what the numbers are actually saying. And still others don't believe that statistics can ever be used in a legitimate manner to point to the truth. But whether it is about significant trends in the population, average salary and unemployment rates, or similarities and differences across stock prices, statistics are an extremely important input to many decisions that we face daily. And understanding how to generate the statistics and interpret them relating to your particular decision can make all the difference between a good decision and a poor one.

For example, suppose that you are trying to sell your house and you need to set a selling price for it. The mean selling price of houses in your area is \$250,000, so you set your price at \$265,000. Perhaps \$250,000 is the price roughly in the middle of several house prices that have ranged from \$200,000 to \$270,000, so you are in the ballpark. However, a mean of \$250,000 could also occur with house prices of \$175,000, \$150,000, \$145,000, \$100,000, and \$780,000. One high price out of five causes the mean to increase dramatically, so you have potentially priced yourself out of most of the market. For this reason, it is important to understand what the term "mean" really represents.

Another compelling reason to understand statistics is that we are living in a quality-driven society. Everything nowadays is related to "improving

quality," a "quality job," or "quality improvement processes." Companies are striving for higher quality in their products and employees and are using such programs as "continuous improvement" and "six-sigma" to achieve and measure this quality. Even the ordinary consumer has heard these terms and needs to understand them in order to be an educated customer

or client. Here again, an understanding of statistics can help you make wise choices related to purchasing behavior.

So as we move from the information age to the knowledge age, it is becoming increasingly important for us to at least understand, if not generate and use, statistics. In this book, Bob Donnelly has done a wonderful job of presenting statistics so that you can improve your ability to look at and comprehend the data you run across every day. Bob's many years of teaching statistics at all levels have provided the basis for his phenomenal ability to explain difficult statistical concepts clearly. Even the most unsophisticated reader will soon understand the subtleties and power of telling the truth with statistics!

Christine T. Kydd

2003 Delaware Professor of the Year

Associate Professor of Business Administration and Director of Undergraduate

Programs

University of Delaware

Introduction

Statistics. Why does this single word terrify so many of today's students? The mere mention of this word in the classroom causes a glassy-eyed, deer-in-the-headlights reaction across a sea of faces. In one form or another, the topic of statistics has been torturing innocent students for hundreds of years. You would think the word statistics had been derived from the Latin words sta, meaning "Why" and tistica, meaning "Do I have to take this %#!\$©*% class?" But it really doesn't have to be this way.

The term "stat" needn't be a four-letter word in the minds of our students.

As you read this paragraph, you're probably wondering what this book can do for you. Well, it's written by a person (that's me) who (a) clearly remembers being in your shoes as a student (even if it was in the last century), (b) sympathizes with your current dilemma (I can feel your pain), and (c) has learned a thing or two over many years of teaching (those many hours of tutorials were not for naught). The result of this experience has allowed me to discover ways to walk you through many of the concepts that traditionally frustrate students. Armed with the tools that you will gain from the many examples and numerous problems explained in detail, this task will not be as daunting as it first appears.

Unfortunately, fancy terms such as inferential statistics, analysis of variance, and hypothesis testing are enough to send many running for the hills. My goal has been to show that these complicated terms are really used to describe ordinary, straightforward concepts. By applying many of the techniques to everyday (and sometimes humorous) examples, I have attempted to show that not only is statistics a topic that anyone can master, but it can also actually make sense and be helpful in numerous situations.

To further help those in need, I have established a companion website for this book at www.stat-guide.com. Here you will find additional problems with solutions and links to other useful websites. If you have any feedback you would like to provide about this book, please send me an e-mail via this website.

So hold on to your hats, we're about to take a wild ride into the realm of numbers, inequalities, and, oh yes, don't forget all those Greek symbols! You will see equations that look like the Chinese alphabet at first glance, but can, in fact, be simplified into plain English. The step-by-step description of each problem will help you break down the process into manageable pieces. As you work the example problems on your own, you will gain confidence and success in your abilities to put numbers to work to provide usable information. And, guess what, that is sometimes how statisticians are born!

How This Book Is Organized

The book is organized into four parts:

In Part 1, "The Basics," we start from the very beginning without any assumptions of prior knowledge. After a brief history lesson to warm you up, we dive into the world of data and learn about the different types of data and the variety of measurement scales that we can use. We also cover how to display data graphically, both manually and with the help of Microsoft Excel. We wrap up Part 1 with learning how to calculate descriptive statistics of a sample, such as the mean and standard deviation.

In Part 2, "Probability Topics," we introduce the scary world of probability theory. Once again, I assume you have no prior knowledge of this topic (or if you did, I assume you buried it in the deep recesses of your brain, hoping to never uncover it). An important topic in this section is learning how to count the number of events, which can really improve your poker skills. After easing you into the basics, we gently slide into probability distributions, such as the normal and binomial. Once you master these, we have set the stage for Part 3.

In Part 3, "Inferential Statistics," we start off learning about sampling procedures and the way samples behave statistically. When these concepts are understood, we start acting like real statisticians by making estimates of populations using confidence intervals. By this time, your own mother wouldn't recognize you! We'll top Part 3 off with a procedure that's near and dear to every statistician's heart—hypothesis testing. With this tool, you can do things like make bold comparisons between the male and female population. I'll leave that one to you.

In Part 4, "Advanced Inferential Statistics," we build on earlier topics and explore analysis of variance, a popular method to compare more than two populations to each other. We will also learn about the chi-square tests, which enable us to determine whether two variables are dependent. And last but not least, we'll discover how simple regression (which, by the way, is not so simple or else it wouldn't be the last topic in the book)

describes the strength and direction of the relationship between two variables. When you're done with these topics, your friends won't believe the words they hear coming from your mouth.

Introduction xix

Extras

Throughout this book, you will come across various sidebars that provide a helping hand when things seem to get a little tough. Many are based on my experience as a teacher with the concepts that I have found to cause students the most difficulty.

def

•t*mi»ticm

These are definitions of statistical jargon explained in a nonthreatening manner, which will help to clarify important concepts. You'll find that their bark is often far worse than their bite.

Random Thoughts

In these sidebars I will give you insights that I find interesting (and hopefully you will, too!) about the current topic. Statistics is full of little-known facts that can help relieve the intensity of the topic at hand.



Bob's Basics.

These are tips and insights Vi that I have accumulated over

the years of helping students master a particular topic. The goal here is to have that light bulb in the brain go off, resulting in the feeling of "I got it!"

_j^\j^, Wrong Number

■* / ^P These are warnings of

-?

potential pitfalls lying in wait for an unsuspecting student to fall into. By taking note of these, you'll avoid the same traps that have ensnared many of your predecessors.

Acknowledgments

There are many people whom I am indebted to for helping me with this project. I'd like to thank Jessica Faust for her guidance and expertise to get me on track in the beginning, Mike Sanders for going easy on me with his initial feedback, and Nancy Lewis, for her valuable opinions during the writing process. I'd also like to thank Mike Thomas and Nancy Wagner for their helpful suggestions with the second edition.

To my colleague and friend, Dr. Patricia Buhler, who introduced me to the publishing industry, convinced me to take on this project, and encouraged me throughout the writing process. This all started with you, Pat.

xx The Complete Idiot's Guide to Statistics, Second Edition

To my in-laws, Lindsay and Marge, who never failed to ask me what chapter I was writing, which motivated me to stay on schedule. Your commitment to each other is a true inspiration for all of us.

To my boss of 10 years at Goldey-Beacom College, Joyce Jones, who rearranged my teaching schedule to accommodate my deadlines. Life at GBC will never be the same after you retire, Joyce. I am really going to miss you. Thank you for your constant support over the years. You have been a great boss and a true friend.

To my friend, Jerry Collarini, who provided many recommendations for changes that appear in this second edition.

To my students who make teaching a pleasure. The lessons that I have learned over the years about teaching were invaluable to me as I wrote this book. Without all of you, I would never have had the opportunity to be an author.

To my children, Christin, Brian, and John, and my stepchildren, Katie, Sam, and Jeff, for your interest in this book and your willingness to let me use your antics as examples in many of the chapters.

And most importantly, to my wife, Debbie, who made this a team effort with all the hours she spent contributing ideas, proofreading manuscripts, editing figures, and giving up family time to help me stay on schedule. Deb's excitement over my opportunity to write this book gave me the courage to accept this challenge. Deb was also the inspiration for many of the examples used in the book, allowing me to share experiences from our wonderful life together. Thank you for your love and your patience with me while writing this book.

Trademarks

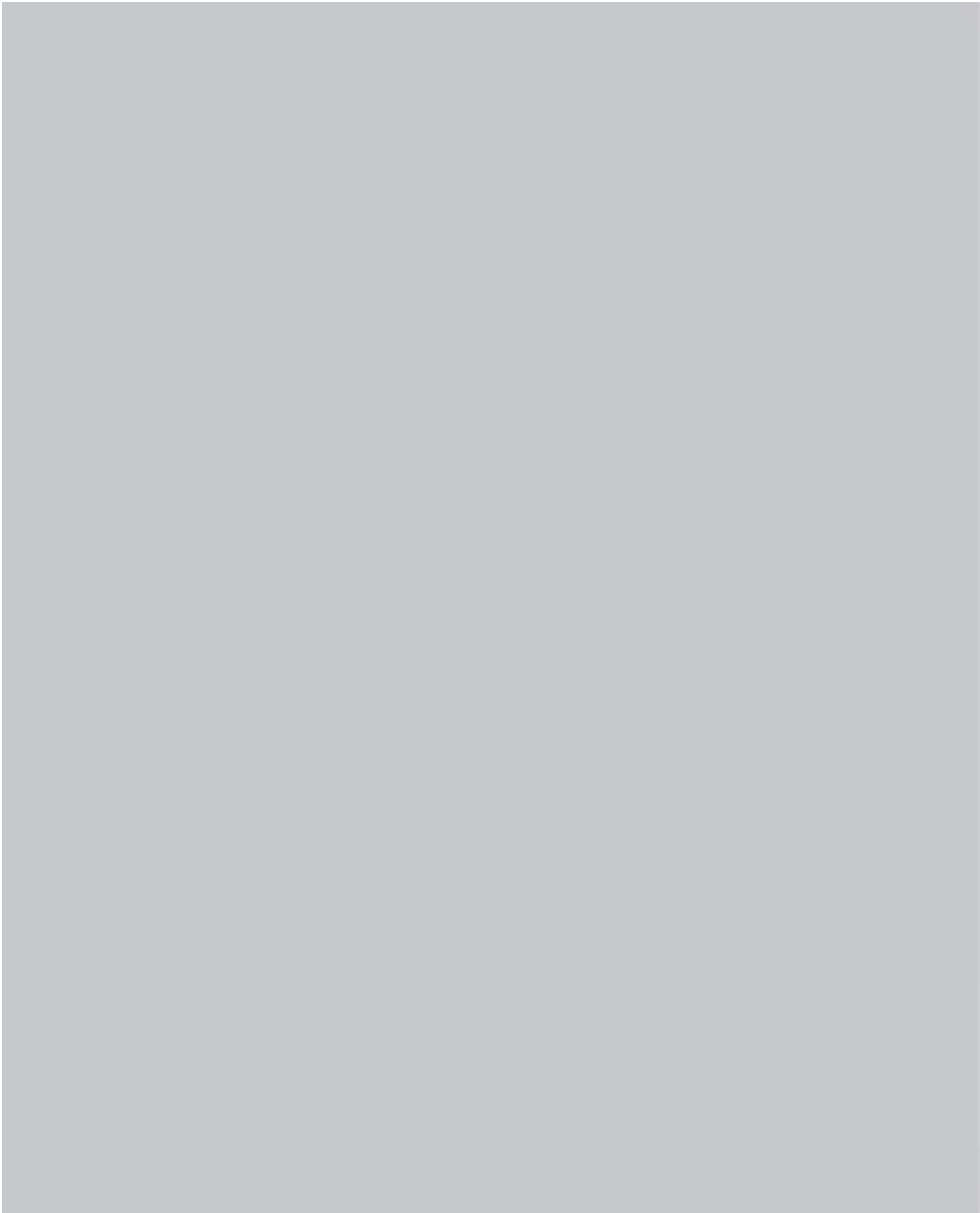
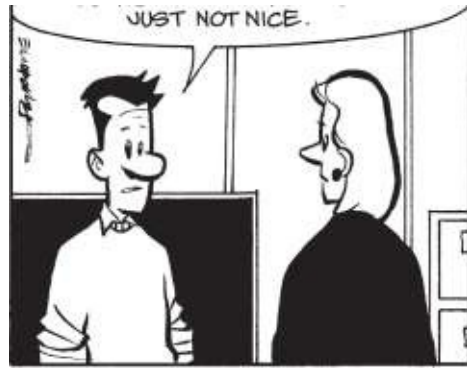
All terms mentioned in this book that are known to be or are suspected of being trademarks or service marks have been appropriately capitalized. Alpha Books and Penguin Group (USA) Inc. cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Part

The Basics

The key to successfully mastering statistics is to have a solid foundation of the basics. To get a firm grasp of the more advanced topics, you need to be well grounded in the concepts presented in this part. After a quick history lesson, these chapters focus on data, the starting point for any method in statistics. You might be surprised with how much there really is to learn about data and all of its properties. We will examine the different types of data, how it is collected, how it is displayed, and how it is used to calculate things called the mean and standard deviation.

THE PIFFEPENCE BETWEEN STANPAEP ANC? MEAN DEVIATION IS THAT ONE IS THE NO£AA ANC7 THE OTHER IS OUST NOT NICE-.



(haptei

Let's Get Started

In This Chapter

- ◆ The purpose of statistics—what's in it for you?
- ◆ The history of statistics—where did this stuff come from?
- ◆ Brief overview of the field of statistics
- ◆ The ethical side of statistics

How many times have you asked yourself why you even need to learn statistics? Well, you're not alone. All too often students find themselves drowning in a mathematical swamp of theories and concepts and never get a chance to see the "big picture" before going under. My goal in this chapter is to provide you with that broader perspective and convince you that statistics is a very useful tool in our current society. In other words, here comes your life preserver. Grab on!

In today's technologically advanced world, we are surrounded by a barrage of data and information from sources trying to convince us to buy something or simply persuade us to agree with their point of view. When we hear on TV that a politician is leading in the polls and in small print see + or - 4 percent, do we know what that means? When a new product is recommended by 4 out of 5 doctors, do we question the validity of the claim? (For instance, were the doctors paid for their endorsement?) Statistics can

Part I: The Basics

have a powerful influence on our feelings, our opinions, and our decisions that we make in life. Getting a handle on this widely used tool is a good thing for all of us.

Where Is This Stuff Used?

tt

The Funk and Wagnalls Dictionary defines statistics as "the science that

deals with the collection, tabulation, and systematic classification of quantitative data, especially as a basis for inference and induction." Now that's a mouthful! In simpler terms, I view statistics as a way to convert numbers into useful information so that good decisions can be made.

These decisions can affect our lives in many ways. For instance, countless medical studies have been performed to determine the effectiveness of new drugs. Statistics form the basis of making an objective decision as to whether this new drug is actually an improvement over current treatments. The results of statistical studies and the manner in which these results are presented often dictate government policies.

Today's corporations are making major business decisions based on statistical analysis. In the 1980s, Marriott conducted an extensive survey with potential customers on their attitudes about current hotel offerings. After analyzing the data, the company launched Courtyard by Marriott, which has been a huge success.

The federal government heavily relies on the national census that is conducted every 10 years to determine funding levels for all the various parts of the country. The statistical analysis performed on this census data has far-reaching implications for many ongoing programs at the state and federal levels.

, Wrong Number

/+* Sp Not interpreting statistical information properly can lead to disaster. Coca-Cola performed a major consumer study in 1985 and, based on the results, decided to reformulate Coke, its flagship drink. After a huge public outcry, Coca-Cola had to backtrack and bring the original formulation back to market. What a mess!

The entire sports industry is completely dependent on the field of statistics. Can you even imagine baseball, football, or basketball without all the statistical analysis that surrounds them? You would never know who the top players are, who is currently hot, and who is in a slump. But then, without statistics, how could the players negotiate those outrageous salaries? Hmmm, maybe I'm onto something here.

My point here is to make you aware of the fact that we are surrounded by

statistics in our society and that our world would be very different if this wasn't the case. Statistics is a useful, and sometimes even critical, tool in our everyday life.

Chapter 1: Let's Get Started

Who Thought of This Stuff?

The field of statistics has been evolving for a very long time. Population surveys appear to be the primary motivation for the historical development of statistics as we know it today. In fact, according to the Bible, Moses conducted a census more than 3,000 years ago. The very word "statistics" comes from the Latin word status, which means "state." This etymological connection reflects the earliest focus of statistics on measuring things such as the number of (taxable) subjects in a kingdom (or state) or the number of subjects to send off to invade neighboring kingdoms.

Early Pioneers

European mathematicians provided the basic foundation for the field of statistics. In 1532, Sir William Petty provided the first accounts of the number of deaths in London on a weekly basis. So began the insurance companies' morbid fascination with death statistics.

During the 1600s, Swiss mathematician James Bernoulli is credited with calculating the probability of a sequence of events, otherwise known as "independent trials." This term is an unfortunate choice of words, as many students over the generations have struggled with this concept and felt like they were on "trial" themselves. You might remember dealing with the problem of calculating (or trying desperately to calculate) the probability of 7 "heads" in 10 coin tosses in a math class. You can thank Mr. Bernoulli for providing you with a way to solve this type of problem. Chapter 9 explores Bernoulli trials in loving detail, and with a little practice you'll get off with a light sentence.

Later, during the 1700s, English mathematician Thomas Bayes developed probability concepts that have also been very useful to the field of statistics. Bayes used the probability of known events of the past to predict probabilities of the future. This concept of inference is widely

used in statistical techniques today. Chapter 7 covers one of his particular contributions, appropriately known as Bayes' theorem.

def'i'ni'tiəm

The term inference refers to a key concept in statistics in which we draw a conclusion from available evidence.

Parti: The Basics

More Recent Famous People

But it wasn't until the early twentieth century that statistics began to develop into the field that we know it as today, when William Gossett developed the famous "t-test" using the Student's t-distribution while working at the Guinness brewery in Dublin, Ireland. We will raise our glasses to Mr. Gossett as we investigate his efforts in Chapter 14.

W. Edwards Deming has been credited with merging the science of statistics with the field of quality control in manufacturing environments. Dr. Deming spent considerable time in Japan during the 1950s and 1960s promoting the concept of statistical quality for businesses. This technique relies on control charts to monitor a process and the use of statistics to determine whether the process is operating satisfactorily. During the 1970s, the Japanese auto industry gained major market share in this country due mainly to superior quality. That's the power of statistics!

Random Thoughts

Dr. Deming's philosophy has been condensed to what is known as Deming's 14 points. This list has proven to be invaluable for organizations seeking ways to use statistics to make their processes more efficient. Through Dr. Deming's efforts, statistics has found a significant role in the business world. Check out his book *The Deming Management Method* (Perigee, 1988) for more information.

The Field of Statistics Today

The science of statistics has evolved into two basic categories known as descriptive statistics and inferential statistics. Because descriptive

statistics is generally simpler, it can be thought of as the "minor league" of the field; whereas inferential statistics, being more challenging, can be considered the "major league" of the two.

def^rnrtn

The purpose of descriptive statistics is to summarize or display data so we can quickly obtain an overview. Inferential statistics allows us to make claims or conclusions about a population based on a sample of data from that population. A population represents all possible outcomes or measurements of interest. A sample is a subset of a population.

Chapter 1: Let's Get Started

Today, computers and software play a dominant role in our use of statistics. Current desktop computers have the capability of processing and analyzing huge amounts of data and information. Specialized software such as SAS and SPSS allows you to conveniently perform all sorts of complicated statistical techniques without breaking a sweat.

In this book, I will show you how to perform many statistical techniques using Microsoft Excel, a spreadsheet software package that's readily available on most desktop computers (also included in the Microsoft Office software suite). Excel has many easy-to-use statistics features that can save you time and energy. If this paragraph causes your blood pressure to elevate (hey, wait a minute, nobody told me this was a computer book!), have no fear. Feel free to just skip over these sections; subsequent material in this book does not depend on this information. I promise it will not be on the final exam.

Descriptive Statistics-the Minor League

The main focus of descriptive statistics is to summarize and display data. Descriptive statistics plays an important role today because of the vast amount of data readily available at our fingertips. With a basic computer and an Internet connection, we can access volumes of data in no time at all. Being able to accurately summarize all of this data to get a look at the "big picture," either graphically or numerically, is the job of descriptive statistics.

There are many examples of descriptive statistics, but the most common is the average. As an example, let's say I would like to get a perspective on the average attention span of my Labrador retriever by using flash cards. I time each incident with a stopwatch and write it down on my clipboard. The following table lists our results, measured in seconds:

Observation	Seconds
-------------	---------

1	4
---	---

2	8
---	---

3	5
---	---

4	10
---	----

5	2
---	---

6	4
---	---

7	7
---	---

8	12
---	----

Parti: The Basics

Using descriptive statistics, I can calculate the average attention span, as follows:

$$4 + 8 + 5 + 10 + 2 + 4 + 7 + 12 + 7 = 66$$
$$= 6.6 \text{ seconds}$$

9

Descriptive statistics can also involve displaying the data graphically, as shown in Figure 1.1. What a good dog!

Figure 1.1

Attention span graph.

</3 "O

C

o o a> CO

1 23456789 Observation

We will delve into descriptive statistics in more detail in Chapters 3 and 4. But until then, we're ready to move up to the big leagues—inferential statistics.

Inferential Statistics—the Major League

As important as descriptive statistics is to us number crunchers, we really get excited about inferential statistics. This category covers a large variety of techniques that allow us to make actual claims about a population based on a sample of data. Suppose, for instance, that I am interested in discovering in general who has the longer attention span, Labrador retrievers or, let's say, teenage boys. (Based on personal observations, I suspect I know the answer to this, but I'll keep it to myself.) Now, it's not possible to measure the attention span of every teenager and every dog, so the next best thing is to take a sample of each and measure them.

At this point, I need to explain the difference between a population and a sample. We use the term "population" in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study. The term "sample" refers to a portion of the population that is representative of the population from which it was

selected.

Chapter 1: Let's Get Started

In this example, the population is all teenage boys and all Labrador retrievers. I need to select a sample of teenagers and a sample of dogs that represent their respective populations. Based on the results of my samples, I can infer the average attention span of each population and determine which is longer.

Figure 1.2 shows the relationship between a population and a sample.

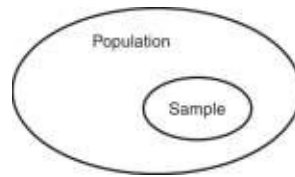


Figure 1.2

The relationship between a population and a sample.

The following are other examples of inferential statistics:

- ◆ Based on a recent sample, I am 95 percent certain that the average age of my customers is between 32 and 35 years old.
- ◆ The average salary for male employees in a particular job category across the country was higher than the female employees' salary, based on a random survey.

In each case, the findings were based on a sample from a larger population and were used to make an inference on that entire population.

The basic difference between descriptive and inferential statistics is that descriptive statistics reports only on the observations at hand and nothing more. Inferential statistics makes a statement about a population based solely on results from a sample taken from that population.

I must tell you at this point that inferential statistics is the area of this field that r, "J\ Bob's Basics . students find the most challenging. To be able to make statements based on samples, you need to use mathematical models that involve probability theory. Now don't panic. Take a deep breath and count to 10 slowly. That's better. I realize that this is often the stumbling block for many, so I have devoted plenty of pages to that nasty "p" word.



A good understanding of probability concepts is an essential stepping-stone for

properly digesting statistics.

Part 2 of this book covers prot

ability.

10 Parti: The Basics

Ethics and Statistics-It's a Dangerous World Out There

People often use statistics when attempting to persuade you to their point of view. Because they are motivated to convince you to purchase something from them or simply to support them, this motivation can lead to the misuse of statistics in several ways.

One of the most common misuses is choosing a sample that ensures results consistent with the desired outcome, rather than choosing a sample representative of the population of interest. This is known as having a biased sample.

. . Suppose, for instance, that I'm an upstanding politi-

—i cian whose only concern is the best interest of my

def

A biased sample is a sample that does not represent the intended population and can lead to distorted findings. Biased sampling can occur either intentionally or unintentionally.

constituents and I want to propose that Congress establish a national golf holiday. During this honored day, all government and business offices would be closed so that we could all run out to chase a little white ball into a hole that's way too small, with sticks purposely designed by the evil golf companies to make this task impossible. Sounds like fun to me! Somehow, I would need to demonstrate that the average level-headed American is in favor of this. Here is where the genius part of my plan lies: rather than survey the general American public, I pass out my survey form only at golf courses. But wait ... it only gets better. I design the survey to look like the following:

We would like to propose a national golf holiday, on which everybody gets the day off from work and plays golf all day. (This means you would not need permission from your spouse.) Are you in favor of this proposal?

A. Yes, most definitely.

B. Sure, why not?

C. No, I would rather spend the entire day at work.

P.S. If you choose C, we will permanently revoke all your golfing privileges everywhere in the country for the rest of your life. We are dead serious.

I can now honestly report back to Congress that the respondents of my survey were overwhelmingly in favor of this new holiday. And from what we know about Congress, they'd probably believe me.

Chapter 1: Let's Get Started 11

Another way to misuse statistics is to make differences seem greater than they actually are by graphically presenting the data in a deceptive manner. Now that I have golf on my brain, let me use my golf scores to demonstrate this point. Let's say, hypothetically speaking of course, that my average golf score during the month of May was 98. After taking some lessons in June, my average score in July dropped to 96. (For you nongolfers, lower is better.) The graph in Figure 1.3 shows that this improvement was nothing to write home about.

120

O o </)

■H

O

CD

CD

cn

CO

i_

CD

100

80

60

40

20

Figure 1.3

This graph shows the actual difference between May and July.

May

July

Month

However, to avoid feeling like I wasted my money on lessons, I can present the difference between May and July on a different scale, as in Figure 1.4.

99

CD

95

May

July

Figure I.A

This graph exaggerates the difference between May and July.

Month

12 Parti: The Basics

By changing the scale of the graph, it appears that I really made progress on my golf game—when in reality, little progress was made. Oh well, back to the drawing board.

Many of the polls we see on the Internet represent another potential misuse of statistics. Many websites encourage visitors to vote on a question of the day. The results of these informal polls are unreliable simply because those collecting the data have no control over who responds or how many times they respond. As stated earlier, a valid statistical study depends on selecting a sample representative from the population of interest. This is not possible when any person surfing the Internet can participate in the poll. Even though most of these polls state that the results are not scientific, it's still a natural human tendency to be influenced by the results we see.

The lesson here is that we are all consumers of statistics. We are constantly surrounded by information provided by someone who is trying to influence us or gain our support. By having a basic understanding about the field of statistics, we increase the likelihood that we can ward off those evil spirits in their attempts to distort the truth. In Chapter 2, we'll begin our journey to achieve this goal ... oh, and to help you pass your statistics course.

Your Turn

Identify each of the following statistics as either descriptive or inferential.

1. Seventy-three percent of Asian American households in the United States own a computer.
2. Households with children under the age of 18 are more likely to have access to the Internet (62 percent) than family households with no

children (53 percent).

3. Hank Aaron hit 755 career home runs.

4. The average SAT score for incoming freshman at a local college was 950.

5. On a recent poll, 67 percent of Americans had a favorable opinion of the President of the United States.

You can find additional sample problems on my website: www.stat-guide.com.

Chapter 1: Let's Get Started 13

The Least You Need to Know

- ◆ Statistics is a vital tool that provides organizations with the necessary information to make good decisions.
- ◆ The field of statistics evolved from the early work of European mathematicians during the seventeenth century.
- ◆ Descriptive statistics focuses on the summary or display of data so we can quickly obtain an overview.
- ◆ Inferential statistics allows us to make claims or conclusions about a population based on a sample of data from that population.
- ◆ We are all consumers of statistics and need to be aware of the potential misuses that can occur in this field.

(hapte



Data, Data Everywhere and Not a Drop to Drink

In This Chapter

- ◆ The difference between data and information
- ◆ Where does data come from?
- ◆ What kinds of data can we use?
- / Different ways of measuring data
- ◆ Setting up Excel for statistical analysis

Data is the basic foundation for the field of statistics. The validity of any statistical study hinges on the validity of the data from the beginning of the process. Many things can come into question, such as the accuracy of the data or the source of the data. Without the proper foundation, your efforts to provide a sound analysis will come tumbling down.

The issues surrounding data can be surprisingly complex. After all, aren't we just talking about numbers here? What could go wrong? Well, plenty can. Because data can be classified in several ways, we need to recognize the difference between quantitative and qualitative data and how each is used.

16 Parti: The Basics

Data also can be measured in many ways. The data measurement choice we make at the start of the study will determine what kind of statistical techniques we can apply.

The Importance of Data

Data is simply defined as the value assigned to a specific observation or measurement. If I'm collecting data on my wife's snoring behavior, I can do so in different ways. I can measure how many times Debbie snores over a 10-minute period. I can measure the length of each snore in seconds. I could also measure how loud each snore is with a descriptive phrase, like "That one sounded like a bear just waking up from hibernation" or "Wow! That one sounded like an Alaskan seal calling for its young." (How a sound like that can come from a person who can fit

into a pair of size 2 jeans and still be able to breathe I'll never know.)

In each case, I'm recording data on the same event in a different form. In the first case, I'm measuring a frequency or number of occurrences. In the second instance, I'm measuring duration or length in time. And the final attempt measures the event by describing volume using words rather than numbers. Each of these cases just shows a different way to use data.

If you haven't noticed yet, statistics people like to use all sorts of jargon, and here are a couple more terms. Data that is used to describe something of interest about a population is called a parameter. However, if the data is describing a sample from that population, we refer to it as a statistic. For instance, let's say that the population of interest is my wife's three-year-old preschool class and my measurement of interest is how many times the little urchins use the bathroom in a day (according to Debbie, much more than should be physically possible).

If we average the number of trips per child, this figure would be considered a parameter because the entire population was measured. However, if we want to make a statement about the average number of bathroom trips per day per three-year-old in the country, then Debbie's class could be our sample. We can consider the average that we observe from her class a statistic if we assume it could be used to estimate all three-year-olds in the country.

Data is the building blocks of all statistical studies. You can hire the most expensive, well-known statisticians and provide them with the latest computer hardware and software available, but if the data you provide them is inaccurate or not relevant to the study, the final results will be worthless.

Chapter 2: Data, Data Everywhere and Not a Drop to Drink 17

However, data all by its lonesome is not all that useful. By definition, data is just the raw facts and figures that pertain to a measurement of interest. Information, on the other hand, is derived from the facts for the purpose of making decisions. One of the major reasons to use statistics is to transform data into information. For example, the table that follows

shows monthly sales data for a small retail store.

def'i'iorticri

Data is the value assigned to an observation or a measurement and is the building blocks to statistical analysis. The plural form is data and the singular form is datum, referring to an individual observation or measurement.

Data that describes a characteristic about a population is known as a parameter. Data that describes a characteristic about a sample is known as a statistic. Information is data that is transformed into useful facts that can be used for a specific purpose, such as making a decision.

Monthly Sales Data

Month Sales (S)

January 15,178

February 14,293

March 13,492

April 12,287

May 11,321

Using statistical analysis, we can generate information that may be of interest, such as "Wake up! You are doing something very wrong. At this rate, you will be out of business by early next year." Based on this valuable information, we can make some important decisions about how to avoid this impending disaster.

The Sources of Data-Where Does All This Stuff Come From?

We classify the sources of data into two broad categories: primary and secondary. Secondary data is data that somebody else has collected and made available for others to use. The U.S. government loves to collect and publish all sorts of interesting data, just in case anyone should need it.

The Department of Commerce handles census

18 Parti: The Basics

def'i'ni'tier

Primary data is data that you have collected for your own use. Secondary data is data collected by someone else that you are "borrowing."

data, and the Department of Labor collects mountains of, you guessed it, labor statistics. The Department of the Interior provides all sorts of data about U.S. resources. For instance, did you know there are 250 species of squirrels in this country? If you don't believe me, go to www.npwrc.usgs.gov/resource/distr/mammals/mammals/_squirrel.htm and you can become the local "squirrel" expert.

The Canadian government has a great system for providing statistical data to the public. Rather than each department in the government being responsible for collecting and disbursing data as in the United States, Canada has a national statistical agency known as Statistics Canada (www.statcan.ca/start.html). It's like one-stop shopping for the statistician. It's a wonderful website that makes research of Canadian facts a pleasure.

The main drawback of using secondary data is that you have no control over how the data was collected. It's a natural human tendency to believe anything that's in print (you believe me, don't you?), and sometimes that requires a leap of faith. The advantage to secondary data is that it's cheap (sometimes free) and it's available now. That's called instant gratification.

Primary data, on the other hand, is data collected by the person who eventually uses this data. It can be expensive to acquire, but the main advantage is that it's your data and you have nobody else to blame but yourself if you make a mess of it.

When collecting primary data, you want to ensure that the results will not be biased by the manner in which it is collected. You can obtain primary data in many ways, such as direct observation, surveys, and experiments.

Random Thoughts

The Internet has also become a rich source of data for statistics published by various industries. If you can muddle your way through the 63,278 sites that come back from the typical Internet search engine, you might find something useful. I once found a Japanese study on the effect of fluoride on toad embryos (www.fluoride-journal.com/_1971.htm). Before this discovery, I was completely oblivious to the fact that toads even had teeth, much less a cavity problem. I can't wait to impress my friends at the next neighborhood dinner party.

Chapter 2: Data, Data Everywhere and Not a Drop to Drink 19

Direct Observation—I'll Be Watching You

Most often, this method focuses on gathering data while the subjects of interest are in their natural environment, oblivious to what is going on around them. Examples of these studies would be observing wild animals stalking their prey in the forest or teenagers at the mall on Friday night (or is that the same example?). The advantage of this method is that the subjects will unlikely be influenced by the data collection.

Focus groups are a direct observational technique where the subjects are aware that data is being collected. Businesses use focus groups to gather information in a group setting controlled by a moderator. The subjects are usually paid for their time and are asked to comment on specific topics.

Experiments-Who's in Control?

This method is more direct than observation because the subjects will participate in an experiment designed to determine the effectiveness of a treatment. An example of a treatment could be the use of a new medical drug. Two groups would be established. The first is the experimental group who receive the new drug, and the second is the control group who think they are getting the new drug but are in fact getting no medication. The reactions from each group are measured and compared to determine whether the new drug was effective.

The claims that the experimental studies are attempting to verify need to be clear and specific. I just recently read about an herb called ginkgo biloba. According to this article, people who make money selling funny-

sounding herbs claim ginkgo biloba will keep your mind sharp as you age. Sounds like something everyone would want. Now let's see, where was I? As stated, this claim might prove difficult to verify. How do you define "keeping your mind sharp"? And then, how do you measure sharpness of mind? These are some of the challenges that statistical experiments face.

The benefit of experiments is that they allow the statistician to control factors that could influence the results, such as gender, age, and education of the participants. The concern about collecting data through experiments is that the response of the subjects might be influenced by the fact that they are participating in a study. The design of experiments for a statistical study is a very complex topic and goes beyond the scope of this book.

20 Parti: The Basics

Surveys-Is That Your Final Answer?

This technique of data collection involves directly asking the subject a series of questions. The questionnaire needs to be carefully designed to avoid any bias (see Chapter 1) or confusion for those participating. Concerns also exist about the influence the survey will have on the participant's responses. Some participants respond in a way they feel the survey would like them to. This is very similar to the manner in which hostages bond with their captors. The survey can be administered by e-mail, snail-mail, or telephone. It's the telephone survey that I'm most fond of, especially when I get the call just as I'm sitting down to dinner, getting into the shower, or finally making some progress on the chapter I'm writing.

Bob's Basics



Research has shown that the manner in which the questions are asked can affect the responses a person provides on a questionnaire. A question posed in a positive tone will tend to invoke a more positive response and vice versa. A good strategy is to test your questionnaire with a small group of people before releasing it to the general public.

Whatever method you employ, your primary concern should always be that the sample is representative of the population in which you are interested.

Types of Data

Another way to classify data is by one of two types: quantitative or qualitative.

♦ Quantitative data uses numerical values to describe something of interest. An example is Debbie's age, which I have been forced by a legally bound document to never, never, never reveal anywhere in this book, not even if it's buried in an appendix as an answer to an obscure question. (Hint: See page 167.)

♦ Qualitative data uses descriptive terms to measure or classify something of interest. One example of qualitative data is the name of a respondent in a survey and his or her level of education. The next section covers qualitative data in more detail.

Chapter 2: Data, Data Everywhere and Not a Drop to Drink 21

Types of Measurement Scales-a Weighty Topic

Who would have thought of so many ways to look at data? The final way to classify data is by the way it is measured. This distinction is critical because it affects which statistical techniques we can use in our analysis of the data. Measurement classification can be made in several levels.

Nominal Level of Measurement

A nominal level of measurement deals strictly with qualitative data. Observations are simply assigned to predetermined categories. One example is gender of the respondent, with the categories being male and female. Another example is data indicating the type of dog, if any, owned by families in a neighborhood. The categories for this data are the various dog types (black Lab, terrier, stupid mangy mutt that keeps me awake by barking all night at the moon). This data type does not allow us to perform any mathematical operations, such as adding or multiplying. We also cannot rank-order this list in any way from highest to lowest

(although I would put black Lab at the top). This type is considered the lowest level of data and, as a result, is the most restrictive when choosing a statistical technique to use for the analysis.

You can use numbers at the nominal level of measurement. Even in this case, the rules of the nominal scale still remain. An example would be zip codes or telephone numbers, which can't be added or placed in a meaningful order of greater than or less than. Even though the data appears to be numbers, it's handled just like qualitative data.

Ordinal Level of Measurement

On the food chain of data, ordinal is the next level up. It has all the properties of nominal data with the added feature that we can rank-order the values from highest to lowest. An example is if you were to have a lawnmower race. Let's say the finishing order was Scott, Tom, and Bob. We still can't perform mathematical operations on this data, but we can say that Scott's lawnmower was faster than Bob's. However, we cannot say how much faster. Ordinal data does not allow us to make measurements between the categories and to say, for instance, that Scott's lawnmower is twice as good as Bob's (it's not).

Ordinal data can be either qualitative or quantitative. An example of quantitative data is rating movies with 1, 2, 3, or 4 stars. However, we still may not claim that a 4-star movie is 4 times as good as a 1-star movie.

22 Parti: The Basics

Interval Level of Measurement

Moving up the scale of data, we find ourselves at the interval level, which is strictly quantitative data. Now we can get to work with the mathematical operations of addition and subtraction when comparing values. For this data, we can measure the difference between the different categories with actual numbers and also provide meaningful information. Temperature measurement in degrees Fahrenheit is a common example here. For instance, 70 degrees is 5 degrees warmer than 65 degrees. However, multiplication and division can't be performed on this data. Why not? Simply because we cannot argue that 100 degrees is twice as warm as 50 degrees.

Ratio Level of Measurement

The king of data types is the ratio level. This is as good as it gets as far as data is concerned. Now we can perform all four mathematical operations to compare values with absolutely no feelings of guilt. Examples of this type of data are age, weight, height, and salary. Ratio data has all the features of interval data with the added benefit of a true 0 point. The term "true zero point" means that a 0 data value indicates the absence of the object being measured. For instance, 0 salary indicates the absence of any salary.

With a true 0 point, we can use the rules of multiplication and division to compare data values. This allows us to say that a person who is 6 feet in height is twice as tall as a 3-foot person or that a 20-year-old person is half the age of a 40 year old.

The distinction between interval and ratio data is a fine line. To help identify the proper scale, use the "twice as much" rule. If the phrase "twice as much" accurately describes the relationship between two values that differ by a multiple of 2, then the data can be considered ratio level.

, Wrong Number

-^P Interval data does not

have a true 0 point. For example, 0 degrees Fahrenheit does not represent the absence of temperature, even though it may feel like it. To help explain this, try baking a cake at twice the recommended temperature in half the recommended time. Yuck!

There are endless examples of ratio data. Let's look at measuring typing speed in words per minute. I happen to be a handicapped, two-finger, hunt-and-peck typist who has tried those darned typing programs more than once and just can't get it. I can type maybe 20 words a minute on a good day. My 15-year-old son, John, on the other hand, is one of those show-offs who types while he's not even looking and can still type 60 words a minute. Because we can correctly say that John types three times faster than me, typing speed is an example of ratio data.

t a Drop to I

Figure 2.1 summarizes the different data scales and how they relate to one another. As we explore different statistical techniques later in this book, we will revisit these different measurement scales. You will discover that specific techniques require certain types of data.

Figure Z.I

Summary of data measurement scales.

Computers to the Rescue

As mentioned in Chapter 1, we will explore the use of Excel in solving some of the statistics problems in this book. If you have no interest in using Excel in this manner, just skip this section. I promise you won't hurt my feelings. The purpose of this last section is to talk about the use of computers with statistics in general and then to make sure your computer is ready to follow us along.

The Role of Computers in Statistics

When I was a youthful engineering undergraduate student during the 1970s, the words "personal computer" had no meaning. I performed calculations on a clever gadget fondly known as a "slide rule." For those of you who weren't even alive during this time period, I've included a picture of one in Figure 2.2.

■ I ' . . ■ ■ ' ■ ■ E

Figure 2.2

Slide rule circa 191\$. (Courtesy of www.hpmuseum.org)

As you can see, this device looks like a ruler on steroids. It can perform all sorts of mathematical calculations but is far from being user friendly. During my freshman year in college, I purchased my first handheld calculator, a Texas Instrument model that could only perform the basic math functions. It was the approximate size of a cash register.

Part I: The Basics

At this point, the only serious statistical analysis was being performed on mainframe computers by people with high levels of programming skill. These people were somewhat "different" from the rest of us. Fortunately, we have advanced from the Dark Ages and now have awesome, user-friendly computing power at our fingertips. Powerful programs such as SAS, SPSS, Minitab, and Excel are readily accessible to those of us who don't know a lick of computer programming and allow us to perform some of the most sophisticated statistical analysis known to humankind.

Parts of this book will demonstrate how to solve some of the statistical techniques using Microsoft Excel. Choosing to skip these parts will not interfere with your grasp of topics in subsequent chapters. This is simply optional material to expose you to statistical analysis on the computer. I also assume you already have a basic working knowledge of how to use Excel.

Installing the Data Analysis Add-In

Our first task is to check whether Excel's data analysis tools are available on your computer. Open the Excel program and left-click with your mouse on the Tools menu as shown in Figure 2.3. From this point on in the book, I'll use the term "click" to mean click the left button on your mouse.

Figure 1.1

Excel's Tools menu.

Notice in the figure that the highlighted item is Data Analysis, which may or may not show up under your Tools menu. If Data Analysis does appear under your Tools menu, skip the rest of this paragraph and the next two and proceed to the following paragraph that begins with "Click on Data Analysis ..."

It's a Drop to I

If Data Analysis does not appear under the Tools menu, you need to add it to the menu. To do so, click on Add-Ins under the same Tools menu. If you don't see Add-Ins under this menu at first, expand the menu items by clicking on the downward arrow at the bottom of the Tools menu list. After clicking on Add-Ins, you should see the dialog box in Figure 2.4.

Random Thoughts

If your Tools menu looks different from the one in Figure 2.3, it might be because all of your available menu items are not currently visible. To make all the menu items visible, click on the Expand symbol at the bottom of the list (the double-downward-pointing arrows).



Analysis ToolPak-VBA P Conditional Sum Wizard P Euro Currency Tools
P Frontline's Mathematical Functions P Internet Assistant VBA P Lookup
Wizard P Solver Add-in P TreePlan [leO'-icTi i'ie~ Add- In

J

Analysis ToolPak Provides function:", and integrate for financial and scientific

Figure U

Excel's Add-Ins dialog box.

This dialog box provides a list of available add-ins for you to use. Click on the empty box for Analysis ToolPak, which places a check mark in it, and then click OK. Now click on the Tools menu again; Data Analysis will now appear in the list.

Random Thoughts

Don't panic if you receive the following message: "Microsoft Excel can't run this add-in. This feature is not currently installed. Would you like to install it now?" If you want to install the Analysis ToolPak, you might need to have your original Microsoft Office CD close by. Click the Yes

button and follow any further instructions. Then, the Data Analysis option will become available on the Tools menu.

26 Parti: The Basics

Click on Data Analysis under the Tools menu to open the dialog box shown in Figure 2.5.

Figure 2.5

Excel's Data Analysis dialog box

Analysis Tools

Anova: Two-Factor With Replication Anova: Two-Factor Without Replication Correlation

t porrcfiti-al Smoothing

F-Test Two-Sample for Variances

Fourier Analysis

Histogram

Moving Average

—

-!

Help |

Your Excel program is now ready to perform all sorts of statistical magic for you as we explore various techniques throughout this book. At this point, you can click Cancel and close out Excel. Each time you open Excel in the future, the Data Analysis option will be available under the Tools menu.

Your Turn

Classify the following data as nominal, ordinal, interval, or ratio. Explain your choice.

1. Average monthly temperature in degrees Fahrenheit for the city of Wilmington throughout the year

2. Average monthly rainfall in inches for the city of Wilmington throughout the year

3. Education level of survey respondents

Level Number of Respondents

High school 168

Bachelor's degree 784

Master's degree 212

4. Marital status of survey respondents

Status Number of Respondents

Single 28

Married 189

Divorced 62

5. Age of the respondents in the survey

Chapter 2: Data, Data Everywhere and Not a Drop to Drink 27

7. The year in which the respondent was born

8. The voting intentions of the respondents in the survey classified as Republican, Democrat, or Undecided

9. The race of the respondents in the survey classified as White, African American, Asian, or Other

10. Performance rating of employees classified as Above Expectations, Meets Expectations, or Below Expectations

11. The uniform number of each member on a sports team

12. A list of the graduating high school seniors by class rank

13. Final exam scores for my statistics class on a scale of 0 to 100

14. The state in which the respondents in a survey reside

The Least You Need to Know

- ◆ Data serves as the building blocks for all statistical analysis.
- ◆ We classify data as either quantitative or qualitative.
- ◆ Nominal data is assigned to categories with no mathematical comparisons between observations.
- ◆ Ordinal data has all the properties of nominal data with the additional capability of arranging the observations in order.
- ◆ Interval data has all the properties of ordinal data with the additional capability of calculating meaningful differences between the observations.
- ◆ Ratio data has all the properties of interval data with the additional capability of expressing one observation as a multiple of another.

Chapter

3

Displaying Descriptive Statistics

In This Chapter

- ◆ How to construct a frequency distribution
- ◆ How to graph a frequency distribution with a histogram
- ◆ How to construct a stem and leaf display
- ◆ The usefulness of pie, bar, and line charts
- ◆ Using Excel's Chart Wizard to construct charts

Having explained the various types of data that exist for statistical analysis in Chapter 2, here we will explore the different ways in which we can present data. In its basic form, making sense of the patterns in the data can be very difficult because our human brains are not very efficient at processing long lists of raw numbers. We do a much better job of absorbing data when it is presented in summarized form through tables and graphs.

In the next several sections, we will examine many ways to present data so that it will be more useful to the person performing the analysis. Through these techniques, we are able to get a better overview of what the data is telling us. And believe me, there is plenty of data out there with some very interesting stories to tell. Stay tuned.

30 Parti: The Basics

Frequency Distributions

One of the most common ways to graphically describe data is through the use of frequency distributions. The best way to describe a frequency distribution is to start with an example.

def

•i # ni # ticn

A frequency distribution is a table that shows the number of data observations that fall into specific intervals.

Ever since I was a young boy, I have been a devoted fan of the Pittsburgh

Pirates Major League Baseball team. Why I still root for these guys, I'll never know, because they have not had a winning season since 1992. Anyway, below is a table of the batting averages of individual Pirates at the end of the 2005 season. I have not attached names with these averages in order to protect their identities.

Pittsburgh Pirates Final Batting Averages for the 2005 Season

It is difficult to grasp what a tough year these guys had by just looking at this data in this table format. Transforming this data into the frequency distribution shown here makes this fact more obvious.

Batting Average

Number of Players

As you can see, a frequency distribution is simply a table that organizes the number of data values into intervals. In this example, the intervals are the batting average ranges in the first column of the table. The number of data values is the number of players

Chapter 3: Displaying Descriptive Statistics il

who fall into each interval shown in the second column. Well, there's always next season to look forward to.

The intervals in a frequency distribution are officially known as classes, and the number of observations in each class is known as class frequencies. Now let's learn how to arrange these classes.

Constructing a Frequency Distribution

You need to make some important decisions when constructing a frequency distribution. To illustrate these decisions, let's use another example, something many of us can relate to—cell phones! My son John and I are on one of those "family share plans," which means he gets all the peak minutes and I get to use my phone between the hours of 3 A.M. and 6 A.M. every other Saturday. The following table represents the number of calls each day during the month of May on John's account.

Calls per Day

Source: A very confusing phone bill that requires a Ph.D. in metaphysical telecommunications to understand.

Using this data, I have constructed the following frequency distribution.

Frequency Distribution

Number of Days

12

8 5 3

1 1

J2 Parti: The Basics

When arranging these classes, I followed these rules:

1. From classes of equal size. I chose 3 data values to be in each class for this distribution. An example of a class is 0-2, which includes the number of days when John made 0, 1, or 2 calls.
2. Make classes mutually exclusive, or in other words, prevent classes from overlapping. For instance, I wouldn't want 2 classes to be 3-5 and 5-7 because 5 calls would be in 2 different classes.
3. Try to have no fewer than 5 classes and no more than 15 classes. Too few or too many classes tend to hide the true characteristics of the frequency distribution.
4. Avoid open-ended classes, if possible (for instance, a highest class of 15-over).

5

defiTi'tien

Classes are considered mutually exclusive when observations can only fall

into one class. For example, the gender classes "male" and "female" are mutually exclusive because a person cannot belong to both classes.

Include all data values from the original table in a class. In other words, the classes should be exhaustive.

Too few or too many classes will obscure patterns in a frequency distribution. Consider the extreme case where there are so many classes that no class has more than one observation. The other extreme is where there is only one class with all the observations residing in that class. This would be a pretty useless frequency distribution!

(A Distant) Relative Frequency Distribution

Another way to display frequency data is by using the relative frequency distribution. Rather than display the number of observations in each class, this method calculates the percentage of observations in each class by dividing the frequency of each class by the total number of observations. I can display John's data as a relative frequency distribution, as I do in the following table.

defi'nrtier

Relative frequency distributions display the percentage of observations of each class

relative to the total number of observations.

Chapter 3: Displaying Descriptive Statistics 33

According to this distribution, John uses his phone 3 to 5 times 2 7 percent of the days during a month.

The total percentage in a relative frequency distribution should be 100 percent or very close (within 1 percent, because of rounding errors).

Cumulative Frequency Distribution

This "kissing cousin" of the relative frequency 1 f # J # M : » ±\r u\

distribution simply totals the percentages of "" each class as you move down the column. (Get it? Cousin, relative? Sorry, I couldn't help myself!) This provides you with the percentage of observations that are less than or equal to the class of interest. The resulting cumulative frequency distribution is shown here.

Cumulative frequency distributions indicate the percentage of observations that are less than or equal to the current class.

Cumulative Frequency Distribution

Vi Parti: The Basics

The value 0.67 in the fourth column is the result of adding 0.40 to 0.27. According to this table, John used his phone 8 times or less on 84 percent of the days in the month.

If designed properly, frequency distribution is an excellent way to tease good information out of stubborn data. Now, let's deal with how to display the distribution graphically.

Graphing a Frequency Distribution-the Histogram

A histogram is simply a bar graph showing the number of observations in each class as the height of each bar. Figure 3.1 shows the histogram for John's phone calls.

def'i'ni'tieio

A histogram is a bar graph showing the number of observations in each class as the height of each bar.

This graph gives us a good visual of John's calling habits. At least the highest class on the graph is the 0 to 2 calls per day. Things could be worse.

Figure 3.1

A histogram of Johns phone calls.

14

w 12

o

•S 10 >

CD 8

co O 6

CD

-Q

E

13

0-2

3-5

6-8 9-11 Calls per Day

12-14 15-17

Letting Excel Do Our Dirty Work

Excel will actually construct a frequency distribution for us and plot the histogram. How nice!

1. The first thing we need to do is open Excel to a blank sheet and enter our data in Column A starting in Cell A1 (use the data from the earlier table).

Chapter 3:

I Descriptive Statistics 35

Next enter the upper limits to each class in Column B starting in Cell B1. For example, in the class 0-2, the upper limit would be 2. Figure 3.2 shows what the spreadsheet should look like.

Random Thoughts

For some bizarre reason, Excel refers to classes as bins. Go figure.

Figure 3.2

Raw data for the frequency distribution.

3. Go to the Tools menu at the top of the Excel window and select Data Analysis. (Refer to the section "Installing the Data Analysis Add-in" from Chapter 2 if you don't see the Data Analysis command on the Tools menu.)

4. Select the Histogram option from the list of Analysis Tools (see Figure 3.3) and click the OK button.

J6 Parti: The Basics

Figure 3.3

Data Analysis dialog box.

Analysis Tools

Anova: Single Factor

Anova: Two-Factor With Replication

Anova: Two-Factor Without Replication

Correlation

Covariance

Descriptive Statistics

Exponential Smoothing

F-Test Two-sample For Variance,

Fourier Analysis

L^J

$*d$

In the Histogram dialog box (as shown in Figure 3.4), click in the Input Range list box, and then click in the worksheet to select cells A1 through A30 (the 30 original data values). Then, click in the Bin Range list box and click in the worksheet to select cells B1 through B6 (the upper limits for the 6 classes).

Click the New Worksheet Ply option button and the Chart Output check box (see Figure 3.4).

Figure 3.4

Histogram dialog box.

7. Click OK to generate the frequency distribution and histogram (see Figure 3.5).

Figure 3.5

Frequency distribution and
histogram.

Chapter 3:

I Descriptive Statistics 37

Notice that Excel has generated the frequency distribution for us in columns A and B. Cool! The problem here is that the histogram looks like an elephant sat on it. Click on the chart to select it, and then click on the bottom border to drag the bottom of the chart down lower, expanding the histogram to look like Figure 3.6.

Frequency distributions and histograms are convenient ways to get an

accurate picture of what your data is trying to tell you. It sounds like my data is telling me to "get more monthly minutes on your cell phone plan." Wonderful.

Random Thoughts

I prefer using Chart Wizard to display the histogram because I think the graph looks better than when I use the Data Analysis tool. The Chart Wizard allows me more control over the final appearance.

Figure 3.6

Final histogram.

Statistical Flower Power-the Stem and Leaf Display

The stem and leaf display is another graphical technique you can use to display your data. A statistician named John Tukey originated the idea during the 1970s. The major benefit of this approach is that all the original data points are visible on the display.

To demonstrate this method, I will use my son Brian's golf scores for his last 24 rounds, shown in the following table. Normally, Brian would only report his better scores, but we statisticians must be unbiased and accurate.

38 Parti: The Basics

Brian's Golf Scores

81 86 78 80

79 83 84 95

84 79 80 83

81 82 92 90

85 88 80 78

79 87 84 80

Figure 3.7 shows the stem and leaf display for these scores.

Figure 3.7

Stem and leaf display.

7|88999

8|00001123344445678 91025

Stem and Leaf Display

def

•i # ni # tien

The "stem" in the display is the first column of numbers, which represents the first digit of the golf scores. The "leaf" in the display is the second digit of the golf scores, with 1 digit for each score. Because there were 5 scores in the 70s, there are 5 digits to the right of 7.

If we choose to, we can break this display down further by adding more stems. Figure 3.8 shows this approach.

Here, the stem labeled 7 (5) stores all the scores between 75 and 79. The stem 8 (0) stores all the scores between 80 and 84. After examining this display, I can see a pattern that's not as obvious when looking at Figure 3.7: Brian usually scores in the low 80s.

You can find an excellent source of information about stem and leaf displays at the Statistics Canada website at www.statcan.ca/english/edu/power/ch8/plots.htm.

The stem and leaf display

splits the data values into stems (the first digit in the value) and leaves (the remaining digits in the value). By listing all of the leaves to the right of each stem, we can graphically describe how the data is distributed.

Chapter 3:

I Descriptive Statistics 39

7(5)188999 8(0)|000011233444

8 (5) 15678 9(0)j 02

9 (5)|5

More Detailed Stem and Leaf Display

Figure 3.8

A more detailed stem and leaf display.

Charting Your Course

Charts are yet another efficient way to summarize and display patterns in a set of data, so let me demonstrate different types of charts that help us "tell it like it is."

What's Your Favorite Pie Chart?

Pie charts are commonly used to describe data from relative frequency distributions. This type of chart is simply a circle divided into portions whose area is equal to the relative frequency distribution. To illustrate the use of pie charts, let's say some anonymous statistics professor submitted the following final grade distribution.

Final Statistics Grades

Grade

Number of Students

Relative Frequency

9 / 3 <

Total = 1.00

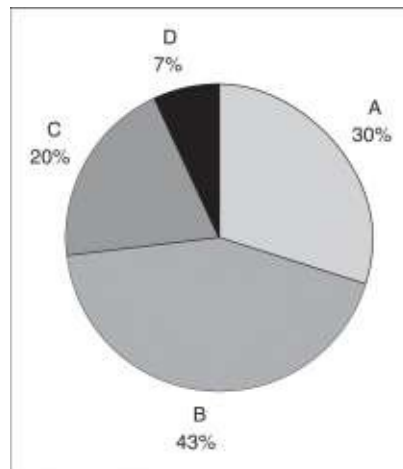
0.30 0.43 0.20 0.07

We can illustrate this relative frequency distribution by using the pie chart in Figure 3.9. This chart was done using Excel's Chart Wizard.

<iO Parti: The Basics

Figure 3.9

Pie chart illustrating a grade distribution.



Bob's Basics



Pie charts are an excellent way to colorfully present data from a relative frequency distribution. If you cannot use colors, use patterns and textures to display pie charts.

As you can see, the pie chart approach is much easier on the eye when compared to looking at data from a table. This person must be a pretty good statistics teacher!

To construct a pie chart by hand, you first need to calculate the center angle for each slice in the pie, which is illustrated in Figure 3.10.

You determine the center angle of each slice by multiplying the relative frequency of the class by 360 (which is the number of degrees in a circle). These results are shown in the following table.

Chapter 3: Displaying Descriptive Statistics 41

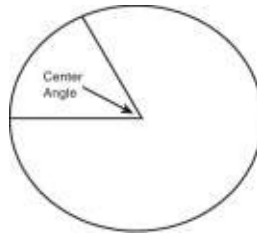


Figure 3.10

The center angle of a pie chart slice.

Center Angle for Pie Chart Construction

Grade Relative Frequency Central Angle

$0.30 * 360 = 108$ degrees $0.43 * 360 = 155$ degrees $0.20 * 360 = 72$ degrees
 $0.07 * 360 = 25$ degrees

By using a device to measure angles, such as a protractor, you can now divide your pie chart into slices of the appropriate size. This assumes, of course, you've mastered the art of drawing circles.

Bar Charts

Bar charts are a useful graphical tool when you are plotting individual data values next to each other. To demonstrate this type of chart (see Figure 3.11), we'll use the data from the following table, which represents the monthly credit card balances for an unnamed spouse of an unnamed person writing a statistics book. (Boy, I'm going to be in big trouble when she sees this.)

U Parti: The Basics

Anonymous Credit Card Balances

Month

Balance (\$)

375 514 834 603 882 468

Source: An unnamed filing cabinet.

Figure 3.11

1000

0

3 4

Month

Random Thoughts

By now you may have just said to yourself, "Hey, wait one minute! Haven't I seen this somewhere before?" By "this" I hope you're referring to the type of chart rather than my wife's credit card statements. The histogram that we visited earlier in the chapter is actually a special type of bar chart that plots frequencies rather than actual data values.

I'm sure your inquisitive mind is now screaming with the question "How do I choose between a pie chart and a bar chart?" If your objective is to compare the relative size of each class to one another, use a pie chart. Bar charts are more useful when you want to highlight the actual data values.

Chapter 3: Displaying Descriptive Statistics W

Line Charts

The last graphical tool discussed here is a line chart (sometimes called a line graph), which is used to help identify patterns between two sets of data. To illustrate the use of line charts, we'll use a favorite topic of mine: teenagers and showers.

Our current resident teenagers seem to have a costly compulsion to take very long, very hot showers, and sometimes more than once a day. As I lie awake at night listening to the constant stream of hot water, all I can envision are dollar bills flowing down the drain. So I have tabulated some

data, which shows the number of showers the cleanest kids on the block have taken in each of the recent months with the corresponding utility bill. Notice that at these rates we average more than one shower per day.

Month

Number of Showers

Utility Bill

72 91

<->H 82 76 85

\$225 \$287 \$260 \$243 \$254 \$275

To see whether there is any pattern between the number of showers and the utility bill, we can plot the pairs of data for each month on a line chart, which is shown in Figure 3.12.

Figure 3.12

A line chart for the number of showers and the utility bill.

70

80 90

Number of Showers

Vi Parti: The Basics

I have chosen to place the number of showers on the x-axis (horizontal) of the chart and the utility bills on the y-axis (vertical) of the chart. Because the line connecting the data points seems to have an overall upward trend, my suspicions hold true. It seems the more showers our waterlogged darlings take, the higher the utility bill.

Line charts prove very useful when you are interested in exploring patterns between two different types of data. They are also helpful when you have many data points and want to show all of them on one graph.

Now that you have mastered the art of displaying descriptive statistics, you are ready to move on to calculating them in the next chapter.

Your Turn

1. The following table represents the exam grades from 36 students from a certain class that I might have taught. Construct a frequency distribution with 9 classes ranging from 56 to 100.
2. Construct a histogram using the solution from Problem 1.
3. Construct a relative and a cumulative frequency distribution from the data in Problem 1.
4. Construct a pie chart from the solution to Problem 1.
5. Construct a stem and leaf diagram from the data in Problem 1 using one stem for the scores in the 50s, 60s, 70s, 80s, and 90s.
6. Construct a stem and leaf diagram from the data in Problem 1 using two stems for the scores in the 50s, 60s, 70s, 80s, and 90s.

Chapter 3: Displaying Descriptive Statistics W

The Least You Need to Know

- ◆ Frequency distributions are an efficient way to summarize data by counting the number of observations in various groupings.
- ◆ Histograms provide a graphical overview of data from frequency distributions.
- ◆ Stem and leaf displays not only provide a graphical display of the data's distribution, but they also contain the actual data values of interest.
- ◆ Pie, bar, and line charts are effective ways to present data in different graphical forms.

Chapter

Calculating Descriptive Statistics: Measures of Central Tendency (Mean,

Median, and Mode)

4

In This Chapter

- ◆ Understanding central tendency
- ◆ Calculating a mean, weighted mean, median, and mode of a sample and population
- ◆ Calculating the mean of a frequency distribution
- ◆ Using Excel to calculate central tendency

The emphasis in Chapter 3 was to demonstrate ways to display our data graphically so that our brain cells could quickly absorb the big picture. With that task behind us, we can now proceed to the next step—summarizing our data numerically. This chapter allows us to throw around some really cool words like "median" and "mode" and, when we're through, you'll actually know what they mean!

Part I: The Basics

As mentioned in Chapter 1, descriptive statistics form the foundation for practically all statistical analysis. If these are not calculated with loving care, our final analysis could be misleading. And as everybody knows, statisticians never want to be misleading. So this chapter focuses on how to calculate descriptive statistics manually and, if you choose, how to verify these results with our good friend Excel.

This is the first chapter that uses mathematical formulas that have all those funny-looking Greek symbols that can make you break out into a cold sweat. But have no fear. We will slay these demons one by one through careful explanation and, in the end, victory will be ours. Onward!

Measures of Central Tendency

There exist two broad categories of descriptive statistics that are commonly used. The first, measures of central tendency, describes the center point of our data set with a single value. It's a valuable tool to help us summarize many pieces of data with one number. The second category, measures of dispersion, is the topic of Chapter 5. But let's explore the many ways to measure the central tendency of our data.

definition

Measures of central tendency describe the center point of a data set with a single value. Measures of dispersion describe how far individual data values have strayed from the mean.

Mean

The most common measure of central tendency is the mean or average, which we calculate by adding all the values in our data set and then dividing this result by the number of observations. The mathematical formula for the mean differs slightly depending on whether you're referring to the sample mean or the population mean. The formula for the sample mean is as follows:

$\bar{x} =$

where:

\bar{x} - the sample mean

x_j - the values in the sample (x_1 = the first data value, x_2 - the second data value, and so on)

$\sum x_i$ = the sum of all the data values in the sample

n =

n = the number of data values in the sample



Don't panic when you see the symbol $\sum_{i=1}^n$, which means "the sum of x , for $i = 1$ to n ." If our data sample contains the values 5, 8, and 2, then $n = 3$, $x_1 = 5$, $x_2 = 8$, and $x_3 = 2$, resulting in the expression:

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 5 + 8 + 2 = 15$$

The formula for the population mean is as follows:

a

* N where:

μ = the population mean (pronounced mil, as in "I hope you find this awz/sing")

77

$\sum x$ = the sum of all the data values in the population

$$\frac{1}{N} = 1$$

N = the number of data values in the population

To demonstrate calculating measures of central tendency, let's use the following example. As in many teenage households, video games are a common form of entertainment in our family room. Brian and John love to challenge me with a game and then clean my clock before I can ask "Which team is mine?" I suspected John of sticking me with the "bad" controller because it felt like a 10-second delay between pushing a button and the game responding. (Turns out the delay was really between my brain and my fingers.) Anyway, the following data set represents the number of hours each week that video games are played in our household.

50 Parti: The Basics

3 7 4 9 5 4 6 17 4

Because this data represents a sample, we will calculate the sample mean:

5> ,

$3 + 7 + 4 + 9 + 5 + 4 + 6 + 17 + 4 + 7 \dots$

= 6.6 hours

n 10

It looks like I need some serious practice time to catch up to these guys.

Weighted Mean

When we calculated the mean number of hours in the previous example, we gave each data value the same weight in the calculation as the other values. A weighted mean refers to a mean that needs to go on a diet. Just kidding; I was checking to see whether you were paying attention. A weighted mean allows you to assign more weight to certain values and less weight to others. For example, let's say your statistics grade this semester will be based on a combination of your final exam score, a homework score, and a final project, each weighted according to the following table.

Type Score Weight (Percent)

50 35 15

We can calculate your final grade using the following formula for a weighted average. Note that here we are dividing by the sum of the weights rather than by the number of data values.

$\sum W_i X_i$

■ 1=1

Bob's Basics

»



The symbol $\sum w_i x_i$ means "the sum of w times x ." Each pair of w and x is

-st

=i

multiplied together, and these results are then summed.

Chapter k: Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 51

where:

W_i = the weight for each data value x_i

77

$\sum w_i$; = the sum of the weights

$\sum_{i=1}^n$

The previous equation can be set up in the following table to demonstrate the procedure.

Type	Score	Weight	Weight x Score
A	94	0.50	47.0
B	89	0.35	31.2
C	83	0.15	12.4
Total		1.00	90.6

$\sum w_i x_i = 90.6$

$\sum_{i=1}^n w_i = 1$

We can obtain the same result by plugging the numbers directly in to the formula for a weighted average:

$\frac{(0.50)(94) + (0.35)(89) + (0.15)(83)}{0.50 + 0.35 + 0.15} = \frac{47.0 + 31.2 + 12.4}{1.0} = 90.6$

$x = 90.6$

$0.50 + 0.35 + 0.15 = 1.0$

Congratulations. You earned an A!

The weights in a weighted average do not need to add to 1 as in the previous example. Let's say I want a weighted average of my two most recent golf scores, 90 and 100, and I want 90 to have twice the weight as 100 in my average. I would calculate this by:

$(2)(90) + (1)(100) = 290 + 100 = 390$ By giving more weight to my lower score, the result is lower than the true average of 95. In this case, I think I'll go with the weighted average.

Mean of Grouped Data from a Frequency Distribution

Here's some great news to get excited about! You can actually calculate the mean of grouped data from a frequency distribution. Recall the data set from Chapter 3 regarding John's cell phone calls per day shown in the following table.

52 Parti: The Basics

Calls per Day

The following table shows this data as a frequency distribution with the calls per day as the class.

Frequency Distribution

To calculate the mean of this distribution, we first need to determine the midpoint of

each class using the following method:

Lower Value + Upper Value Class Midpoint =

For instance, the class midpoint for the last class would be as follows:

Class Midpoint = 16

We can use the following table to assist in the calculations.

Chapter k: Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 5J

Class Midpoint (x) Frequency (/)

12

After we have determined the midpoint for each class, we can calculate the mean of the frequency distribution using the following equation—which is basically a weighted average formula:

$\sum_{i=1}^n$

$x =$

$\frac{77}{7} = 11$

where:

x_i = the midpoint for each class for $i = 1$ to n

f_i = the number of observations (frequency) of each class for $i = 1$ to n

n = the number of classes in the distribution

We determine the mean of this frequency distribution as follows $\frac{1}{n} \sum_{i=1}^n x_i f_i$
 $= \frac{1}{16} [(8)(4) + (5)(7) + (3)(10) + (1)(13) + (1)(16)]$

$= \frac{12+8+5+3+1+1}{16}$

4.6 calls

According to the mean of this frequency distribution, John averages 4.6 calls per day on his cell phone.

$\sum_{j=1}^n u_j$, Wrong Number

$\frac{1}{n} \sum_{i=1}^n x_i f_i$ The mean of a free

The mean of a frequency distribution where data is grouped into classes is only an approximation to the mean of the original data set from which it was derived. This is true because we make the assumption that the original data values are at the midpoint of each class, which is not necessarily the case. The true mean of the 30 original data values in the

cell phone example is only 4.5 calls per day rather than 4.6.

5*1 Parti: The Basics

If the classes in the frequency distribution are a single value rather than an interval, calculate the mean by treating the distribution as a weighted mean. For example, let's say the following table represents the number of days that a hardware store experienced various daily demands for a particular hammer during the past 65 days of business.

Daily Demand (x) Number of Days (f)

0 10

1 15

2 12

3 18

4 6

5 4

Total = 65

For instance, there were 15 days in the past 65 days that the store experienced demand for one hammer. What is the average daily demand during the past 65 days?

* = ■ * = !

If, ; = l

___ (l0)(0) + (l5)(l) + (l2)(2) + (l8)(3) + (6)(4) + (4)(5)

10 + 15 + 12 + 18 + 6 + 4

x = = 2.1 hammers per day

Now that we have become experts in every conceivable method for

calculating a mean, we are ready to move on to the other cool methods of measuring central tendency.

Median

Another way to measure central tendency is by finding the median. The median is the value in the data set for which half the observations are higher and half the observations are lower. We find the median by arranging the data values in ascending order and identifying the halfway point.

Chapter k: Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 55

Using our example with the video games, we rearrange our data set in ascending order:

3

7 7

17

Because we have an even number of data points (10), the median is the average of the two center points. In this case, that will be the values 5 and 6, resulting in a median of 5.5 hours of video games per week. Notice that there are four data values to the left (3, 4, 4, and 4) of these center points and four data values to the right (7, 7, 9, and 17).

To illustrate the median for a data set with an odd number of values, let's remove 17 from the video games data and repeat our analysis.

def

•riortiein

The median is a measure of central tendency that represents the value in the data set for which half the observations are higher and half the observations are lower. When there is an even number of data points, the median will be the average of the two center points.

3

7 7

In this instance, we only have one center point, which is the value 5. Therefore, the median for this data set is 5 hours of video games per week. Again, there are four data values to the left and right of this center point.

Mode

The last measure of central tendency on my mind is the mode, which is simply the observation in the data set that occurs the most frequently.

To illustrate the mode for a data set, let's again use the original video game data.

3 4 4 4 5 6 7 7 9 17

The mode is 4 hours per week because this value occurs three times in the data set.

That wraps up all the different ways to measure central tendency of our data set. However, one question is screaming to be answered, and that

Random Thoughts

There can be more than one mode of a data set if more than one value occurs the most frequent number of times.

56 Parti: The Basics

How Does One Choose?

I bet you never thought you would have so many choices of measuring central tendency! It's kind of like being in an ice cream store in front of 30 flavors. If you think all the data in your data set is relevant, then the mean is your best choice. This measurement is affected by both the number and magnitude of your values. However, very small or very large values can have a significant impact on the mean, especially if the size of

the sample is small. If this is a concern, perhaps you should consider using the median. The median is not as sensitive to a very large or small value.

Consider the following data set from the original video game example:

3 4 4 4 5 6 7 7 9 17

The number 17 is rather large when compared to the rest of the data. The mean of this sample was 6.6, whereas the median was 5.5. If you think 17 is not a typical value that you would expect in this data set, the median would be your best choice for central tendency.

The poor lonely mode has limited applications. It is primarily used to describe data at the nominal scale—that is, data that is grouped in descriptive categories such as gender. If 60 percent of our survey respondents were male, then the mode of our data would be male.

Using Excel to Calculate Central Tendency

Excel will kindly calculate the mean, median, and mode for you all at once with a few mouse clicks. Let's demonstrate this using the data set from the video game example.

1. To begin, open a blank Excel worksheet and enter the video game data (Figure 4.1).
2. Click on the Tools menu at the top of the spreadsheet (between Format and Data) and select Data Analysis. (See the section "Installing the Data Analysis Add-in" in Chapter 2 for more details on this step if you don't see the Data Analysis command.) After selecting Data Analysis, you should see the dialog box shown in Figure 4.2.

Chapter k: Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 57

Figure A.1

Enter data from the video game example.

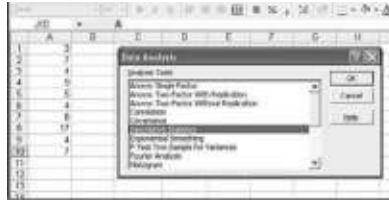


Figure 4.2

Data Analysis dialog box.

3. Select Descriptive Statistics and click OK. The following dialog box will appear (Figure 4.3).

Figure 4.3

Descriptive Statistics dialog box.

58 Parti: The Basics

4. For the Input Range, select cells A1 through A10, select the Output Range option, and select cell C1. Then choose the Summary statistics check box and click OK.

5. After you expand columns C and D slightly to see all the figures, your spreadsheet should look like Figure 4.4.

Figure k.k

Measures of central tendency for the video game example.

As you can see, the mean is 6.6 hours, the median is 5.5 hours, and the mode is 4.0 hours. Piece of cake!

Your Turn

1. Calculate the mean, median, and mode for the following data set: 20, 15, 24, 10, 8, 19, 24, 12, 21, 6.

2. Calculate the mean, median, and mode for the following data set: 84, 82, 90, 77,

75, 77, 82, 86, 82.

3. Calculate the mean, median, and mode for the following data set: 36, 27, 50, 42, 27, 36, 25, 40, 29, 15.

4. Calculate the mean, median, and mode for the following data set: 8, 11, 6, 2, 11, 6, 5, 6, 10.

Chapter k: Calculating Descriptive Statistics: Measures of Central Tendency (Mean, Median, and Mode) 59

5. A company counted the number of their employees in each of the following age classes. According to this distribution, what is the average age of the employees in the company?

Number of Employees

6. Calculate the weighted mean of the following values with the corresponding weights.

Value Weight

118 3

125 2

107 1

7. A company counted the number of employees at each level of years of service in the following table. What is the average number of years of service in this company?

60 Parti: The Basics

The Least You Need to Know

◆ Calculate the mean of a data set by summing all the values and dividing this result by the number of values.

◆ The median of a data set is the midpoint of the set if the values are arranged in ascending or descending order.

♦ The median is the single center value from the data set if there is an odd number of values in the set. The median is the average of the two center values if the number of values in the set is even.

♦ The mode of a data set is the value that appears most often in the set. There can be more than one mode in a data set.

Chapter

Calculating Descriptive Statistics: Measures of Dispersion

5

In This Chapter

- ♦ Calculating the range of a sample
- ♦ Calculating the variance and standard deviation of a sample and population
- ♦ Using the empirical rule and Chebyshev's theorem to predict the distribution of data values
- ♦ Using measures of relative position to identify outlier data values
- ♦ Using Excel to calculate measures of dispersion

In Chapter 4, we calculated measures of central tendency by summarizing our data set into a single value. But in doing so, we lost information that could be useful. For the video game example, if the only information I provided you was that the mean of my sample was 6.6 hours, you would not know whether all the values were between 6 and 7 hours or whether the

62 Parti: The Basics

values varied between 1 and 12 hours. As you will see later, this

distinction can be very important.

To address this issue, we rely on the second major category of descriptive statistics, measures of dispersion, which describes how far the individual data values have strayed from the mean. So let's look at the different ways we can measure dispersion.

Range

The range is the simplest measure of dispersion and is calculated by finding the difference between the highest value and the lowest value in the data set. To demonstrate how to calculate the range, I'll use the following example.

One of Debbie's special qualities is that she is a dedicated grill-a-holic when it comes to barbequing in the backyard. Recently, we needed to purchase a new grill since our old one mysteriously caught fire one night when I was at school teaching. The cause of the fire was labeled "suspicious" after Debbie saw this event as a wonderful opportunity to "upgrade." My idea of the perfect grill is one you dump charcoal in, add one can of lighter fluid to, toss in a match, and run for your life. The best part about this kind of grill is that it has about four parts to assemble, which is something I can easily put together in 3 or 4 weeks. Within minutes of arriving at the store, I felt a stabbing pain in my chest upon finding my wife in an animated conversation with a

total stranger about a grill for the "serious barbeque I r • i • person" complete with three burners, electronic ieni-

aef # i # ni # tien

Obtain the range of a sample

by subtracting the smallest measurement from the largest measurement.

tion, back-up propane tank, 300 horsepower, and front disc brakes. This thing could barbeque a pig on a spit faster than I could say "oink." I'd have better luck assembling a car from scratch. As protection from future acts of arson, I purchased a life insurance policy for our new family member.

Anyway, the following data set represents the number of meals each month that Debbie cranks up on the turbo-charged grill:

7 9 8 11 4

The range of this sample would be: $\text{Range} = 11 - 4 = 7$ meals

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 63

The range is a "quick-and-dirty" way to get a feel for the spread of the data set. However, the limitation is that it only relies on two data points to describe the variation in the sample. No other values between the highest and lowest points are part of the range calculation.

Variance

One of the most common measurements of dispersion in statistics is the variance, which summarizes the squared deviation of each data value from the mean. The formula for the sample variance is:

-1

where:

def'i'ni'tiem

.r = the variance of the sample

X = the sample mean

n = the size of the sample (the number of data values)

$(X_j - \bar{x})$ = thg deviation from the mean for each value in the data set

The variance is a measure of dispersion that describes the relative distance between the data points in the set and the mean of the data set. This measure is widely used in inferential statistics.

The first step in calculating the variance is to determine the mean of the data set, which in the grilling example is 7.8 meals per month. The rest of the calculations can be facilitated by the following table.

7

9

8

11

4

-0.8 1.2 0.2 3.2

-3.8

0.64

1.44

0.04

10.24

6k Parti: The Basics

$\sum (X_i - \bar{X})^2 = 26.80$

$n - 1 = 5$

The final sample variance calculation becomes this:

$s^2 = 5.36$

For those of us who like to do things in one step, we can also do the entire variance

calculation in the following equation:

$$s^2 = \frac{1}{n} [(7-7.8)^2 + (9-7.8)^2 + (8-7.8)^2 + (11-7.8)^2 + (4-7.8)^2]$$

$s^2 = 5.36$

5-1

Using the Raw Score Method (When Grilling)

A more efficient way to calculate the variance of a data set is known as the raw score method. Even though at first glance this equation may look more imposing, its bark is much worse than its bite. Check it out and decide for yourself what works best for you.

$\sum (X - \bar{X})^2$

$\div n$

$n - 1$

where:

$\sum X^2$ = the sum of each data value after it has been squared

$n \times \bar{X}^2$

$\sum X^2$ = the square of the sum of all the data values $\sum X$ Okay, don't have heart failure just yet. Let me lay this out in the following table to prove to you there are fewer calculations than with the previous method.

7 49 9 81

8 64 11 121

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 65

$\sum (X - \bar{X})^2$

$\div n$

$\sum X^2 - n(\bar{X})^2 = 1521$

$331 -$

1521

331-304.2

6.7

As you can see, the results are the same regardless of the method used. The benefits of the raw score method become more obvious as the size of the sample (n) gets larger.

Bob's Basics



\» If you are calculating the variance by hand, my advice is to do your fingers and calculator battery a favor and use the raw score method.

The Variance of a Population

So far, we have discussed the variance in the context of samples. The good news is the variance of a population is calculated in the same manner as the sample variance. The bad news is I need to introduce another funny-looking Greek symbol: sigma. The

equation for the variance of a population is as follows:

n

N

where:

σ^2 = the variance of the population (pronounced "sigma squared")

X_i - the measurement of each item in the population

\bar{X} = the population mean

N = the size of the population

Wrong Number

-J

Notice that the denominator for the population variance equation is N, whereas the denominator for the sample variance is $n - 1$.

66 Parti: The Basics

The raw score version of this equation is:

N

,2 _ «=1

(N I'

N

N Even though this procedure is identical to the sample variance, let me demonstrate with another example. Let's say I am considering my statistics class as my population and the following ages are the measurement of interest. (Can you guess which one is me? My age adds a little spice to the variance.)

21 23 28 47 20 19 25 23

I'll use the raw score method for this calculation with the population size (N) equal to 8. (I'd love to see a class this size.)

21 23 28 47 20 19 25 25

441 529 784 2209 400 361 625 529

]>V = 206

JV = 5878

£#. =(206) 2 = 42436 42436

5878-

$$25878 - 5304.5 = 71.7$$

Thanks to the old guy in the class, the population variance is 71.7.

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 67

Standard Deviation

This method is pretty straightforward. The standard deviation is simply the square root of the variance. Just as with the variance, there is a standard deviation for both the sample and population, as shown in the following equations.

Sample standard deviation:

$$s = \sqrt{v}$$

$$\sigma = \sqrt{V}$$

Population standard deviation:

$$s = \sqrt{v}$$

A standard deviation is the square root of a variance.

$$\sigma = \sqrt{V}$$

To calculate the standard deviation, you must first calculate the variance and then take the square root of the result. Recall from the previous sections that the variance from my sample of the number of meals Debbie grilled per month was 6.7. The standard deviation of this sample is as follows:

$$s = \sqrt{v} = \sqrt{6.7} = 2.6 \text{ meals}$$

Also recall the variance for the age of my class was 71.7. The standard deviation of the age of this population is as follows:

$$\sigma = \sqrt{V} = \sqrt{71.7} = 8.5 \text{ years}$$

The standard deviation is actually a more useful measure than the variance because the standard deviation is in the units of the original data set. In comparison, the units of the variance for the grill example would be 6.7 "meals squared," and the units of the variance for the age example would be 71.7 "years squared." I don't know about you, but I'm not too thrilled having my age reported as 2,209 squared years. I'll take the standard deviation over the variance any day.

Calculating the Standard Deviation of Grouped Data

The following equation shows how to calculate the standard deviation of grouped data in a frequency distribution.

68 Parti: The Basics

$$s = \sqrt{\frac{\sum (X_j - \bar{X})^2 f_j}{n}}$$

s-

n-\ where:

/ = the number of data values in each frequency class

m = the number of classes

n = $\sum f_j$ - the total number of values in the data set

; = 1

The following table is a frequency distribution that represents the number of times each child in Debbie's 3-year-old preschool class needs a "potty break" in a day.

Number of Potty Breaks per Day (X) Number of Children (f)

2 1

3 4

4 12

5 8

6 5

In this example, $m = 5$ and $n = 30$. From Chapter 4, we know the mean of this frequency distribution is this:

$X(M)$

$\sum_{i=1}^m x_i f_i = 31.20$

in

If,

$\bar{x} = 4.4$

$\sum_{i=1}^m x_i f_i = 1(2) + (4)(3) + (12)(4) + (8)(5) + (5)(6) = 31.20$

: 4.4 times per child per day

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 69

$\sum_{i=1}^m x_i f_i = 31.20$

$\sum_{i=1}^m x_i^2 f_i = 139.20$

$s^2 = 1.08$

$s = 1.04$

/1.08 = 1.04 times per child per day

The potty break frequency distribution has a mean of 4.4 times per child per day and a standard deviation of 1.04 times per child per day. The frequency of these potty breaks must keep Debbie very busy.

The Empirical Rule: Working the Standard Deviation

The values of many large data sets tend to cluster around the mean or median so that the data distribution in the histogram resembles a bell-

shape, symmetrical curve. When this is the case, the empirical rule (sounds like a decree from the emperor) tells us that approximately 68 percent of the data values will be within one standard deviation from the mean.

def

•riortiein

For example, suppose that the average exam score for my large statistics class is 88 points and the standard deviation is 4.0 points and that the distribution of grades is bell-shape around the mean, as shown in Figure 5.1. Because one standard deviation above the mean would be 92 ($88 + 4$) and one standard deviation below the mean would be 84 ($88 - 4$), the empirical rule tells me that approximately 68 percent of the exam scores will fall between 84 and 92 points.

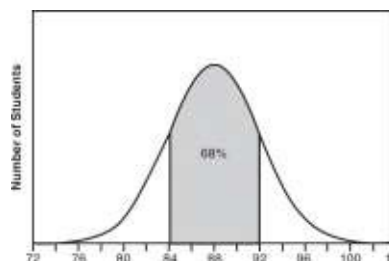
According to the empirical rule,

if a distribution follows a bell-shape—a symmetrical curve centered around the mean—we would expect approximately 68, 95, and 99.7 percent of the values to fall within one, two, and three standard deviations around the mean respectively.

70 Parti: The Basics

Figure 5.1

One standard deviation around the mean.



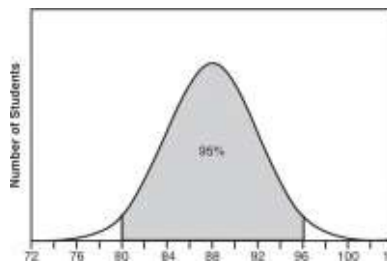
84 88 92

Exam Scores

The empirical rule also states that approximately 95 percent of the data values will fall within two standard deviations from the mean. In our example, two standard deviations equal 8.0 points ($2 * 4.0$). Two standard deviations above the mean would be a score of 96 ($88 + 8$), and two standard deviations below the mean would be 80 ($88 - 8$). According to Figure 5.2, approximately 95 percent of the exam scores will be between 80 and 96 points.

Figure 5.2

Two standard deviations around the mean.



84 88 92

Exam Scores

Taking this one final step, the empirical rule states that, under these conditions, approximately 99.7 percent of the data values will fall within three standard deviations from the mean. According to Figure 5.3, virtually all the test scores should fall within plus or minus 12 points ($3 * 4.0$) from the mean of 88. In this case, I would expect all the exam scores to be between 76 and 100.

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 71

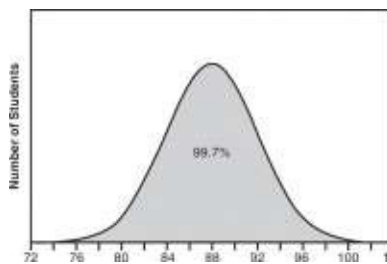


Figure 5.3

Three standard deviations around the mean.

84 88 92

Exam Scores

In general, we can use the following equation to express the range of values within k standard deviations around the mean:

$\bar{x} \pm k\sigma$ We will revisit the empirical rule concept in subsequent chapters.

Chebyshev's Theorem

Chebyshev's theorem is a mathematical rule similar to the empirical rule except that it applies to any distribution rather than just bell-shape, symmetrical distributions. Chebyshev's theorem states that for any number k greater than 1, at least

$\frac{100}{k^2 - 1}$ percent of the values will fall within k standard deviations from the

mean. Using this equation, we can state the following:

At least 75 percent of the data values will fall within two standard deviations from the mean by setting $k = 2$ into Chebyshev's equation.

At least 88.9 percent of the data values will fall within three standard deviations from the mean by setting $k = 3$ into the equation.

Wrong Number

_*

Chebyshev's theorem can be applied to any distribution of data but can only be stated for values of k that are greater than 1 .

72 Parti: The Basics

At least 93.7 percent of the data values will fall within four standard deviations from the mean by setting $k = 4$ into the equation. This last example is shown as:

1

$$\times 100\% = 93.7\%$$

Let's check out Chebyshev's theorem to see whether it really works. The following table shows the number of home runs hit by the top 40 players in Major League Baseball during the 2002 season.

Number of Home Runs from Top 40 Players in 2002

The following histogram shows that this distribution is neither bell-shaped nor symmetrical, so we cannot apply the empirical rule (see Figure 5.4) but will need to use Chebyshev's theorem.

Figure U

Home run histogram for 2002 season.

01

18 16 14 12 10

23-27 28-32 33-37 38-42 43-47 48-52 53-57 Home Runs

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 73

The mean for this distribution is 34.7 home runs, and the standard deviation is 7.2 home runs. The following table summarizes various intervals around the mean with the percentage of values within those intervals.

This table supports Chebyshev's theorem, which predicts that at least 75 percent of the values will fall within two standard deviations from the mean. From the data set, we can observe that 95 percent actually fall between 20.3 and 49.1 home runs (38 out of 40). The same explanation holds true for three and four standard deviations around the mean.

Measures of Relative Position

Another way of looking at dispersion of data is through measures of

relative position, which describe the percentage of the data below a certain point. This technique includes quartile and interquartile measurements.

Quartiles

Quartiles divide the data set into four equal segments after it has been arranged in ascending order. Approximately 25 percent of the data points will fall below the first quartile, Q_1 . Approximately 50 percent of the data points will fall below the second quartile, Q_2 . And, you guessed it, 75 percent should fall below the third quartile, Q_3 . To demonstrate how to identify Q_1 , Q_2 , and Q_3 , let's use the following data set.

def

•nrtien

Quartiles measure the relative position of the data values by dividing the data set into four equal segments.

10 14 6

12 7

ft Parti: The Basics

Step 1: Arrange your data in ascending order.

3 5 6 7 9 10 12 14 14

Step 2: Find the median of the data set. This is Q_2 . 3 5 6 7 9 10 12 14 14

$Q_2 = 9$

Step 3: Find the median of the lower half of the data set (in parenthesis). This is Q_1 . (3 5 6 7) 9 10 12 14 14

$Q_1 = 5.5$ $Q_2 = 9$

Step 4: Find the median of the upper half of the data set (in parenthesis). This is Q_3 . 3 5 6 7 9 (10 12 14 14)

$$Q_1 = 5.5$$

$$Q_2 = 9 \quad Q_3 = 13$$

Interquartile Range

When you have established the quartiles, you can easily calculate the interquartile range (IQR); the IQR measures the spread of the center half of our data set. It is simply the difference between the third and first quartiles, as follows:

def

•rmrticn

The interquartile range measures the spread of the center half of the data set and identifies outliers, which are extreme values that you should discard before analysis.

$$IQR = Q_3 - Q_1$$

The interquartile range is used to identify outliers, which are the "black sheep" of our data set. These are extreme values whose accuracy is questioned and can cause unwanted distortions in statistical results. Any values that are more than:

$$Q_3 + 1.5IQR$$

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 75

or less than:

$$Q_1 - 1.5IQR$$

should be discarded. Let's use the following data set to determine if any nasty outliers exist:

10 42 45 46 51 52 58 73

Since there are eight data values, Q_2 will be the median of the first four values (the midpoint between the second and third values).

$$Q_1 = \hat{Q}_1 = 43,$$

Likewise, Q_3 will be the median of the last four values (the midpoint between the sixth and seventh values).

$$Q_3 = \hat{Q}_3 = 56$$

$$IQR = Q_3 - Q_1 = 56 - 43.5 = 12.5$$
 Any values greater than

$$Q_3 + 1.5IQR = 56 + (1.5)(12.5) = 74.75$$
 or less than

$$Q_1 - 1.5IQR = 43.5 - (1.5)(12.5) = 24.75$$

would be considered an outlier. Therefore, the value 10 would be an outlier in this data set.

Now that we have worked our fingers to the bones calculating all this stuff, let's see how Excel makes it look so easy.

Using Excel to Calculate Measures of Dispersion

Excel enables you to conveniently calculate the range, variance, and standard deviation of a sample using the Data Analysis selection under the Tools menu. Use the exact same steps to calculate these measures as those used to calculate measures of central tendency shown in Chapter 4. Repeating those steps (see the section "Using Excel to Calculate Central Tendency") with the grilling example from this chapter will produce Figure 5.5.

76 Parti: The Basics

Figure 5.5

Measures of dispersion for the turbo grill example.

As you can see from Figure 5.5, the sample range equals 7 meals, the sample variance equals 6.7, and the standard deviation equals 2.6 meals. Also note that this data set has no mode since no value appears more than once.

This wraps up our discussion on the different ways to describe measures of dispersion.

$\sum (x_j - \bar{y})^2$ Wrong Number

$\frac{1}{n}$ Sp The values for variance and standard deviation reported by Excel are for a sample. If your data set represents a population, you need to recalculate the results using N in the denominator rather than $n - 1$.

Your Turn

1. Calculate the variance, standard deviation, and the range for the following sample data set: 20, 15, 24, 10, 8, 19, 24.
2. Calculate the variance, standard deviation, and the range for the following population data set: 84, 82, 90, 77, 75, 77, 82, 86, 82.
3. Calculate the variance, standard deviation, and the range for the following sample data set: 36, 27, 50, 42, 27, 36, 25, 40.
4. Calculate the quartiles and the cutoffs for the outliers for the following data set: 8, 11, 6, 2, 11, 6, 5, 6, 10, 15.
5. A company counted the number of their employees in each of the age classes as follows. According to this distribution, what is the standard deviation for the age

Chapter 5: Calculating Descriptive Statistics: Measures of Dispersion 77

of the employees in the company?

Age Range Number of Employees

20-24 8

25-29 37

30-34 25

35-39 48

40-44 27

45-49 10

6. A company counted the number of employees at each level of years of service in the table that follows. What is the standard deviation for the number of years of service in this company?

Years of Service Number of Employees

1 5

3 10

4 8

5 12

6 3

7. A data set that follows a bell-shape and symmetrical distribution has a mean equal to 75 and a standard deviation equal to 10. What range of values centered around the mean would represent 95 percent of the data points?

8. A data set that is not bell-shape and symmetrical has a mean equal to 50 and a standard deviation equal to 6. What is the minimum percent of values that would fall between 38 and 62?

78 Parti: The Basics

The Least You Need to Know

♦ The range of a data set is the difference between the largest value and smallest value.

♦ The variance of a data set summarizes the squared deviation of each data value from the mean.

♦ The standard deviation of a data set is the square root of the variance and is expressed in the same units as the original data values.

- ♦ The empirical rule states that if a distribution follows a bell-shape, a symmetrical curve centered around the mean, we would expect approximately 68, 95, and 99.7 percent of the values to fall within one, two, and three standard deviations around the mean, respectively.
- ♦ The interquartile range measures the spread of the center half of the data set and identifies outliers, which are extreme values that you should discard before analysis.

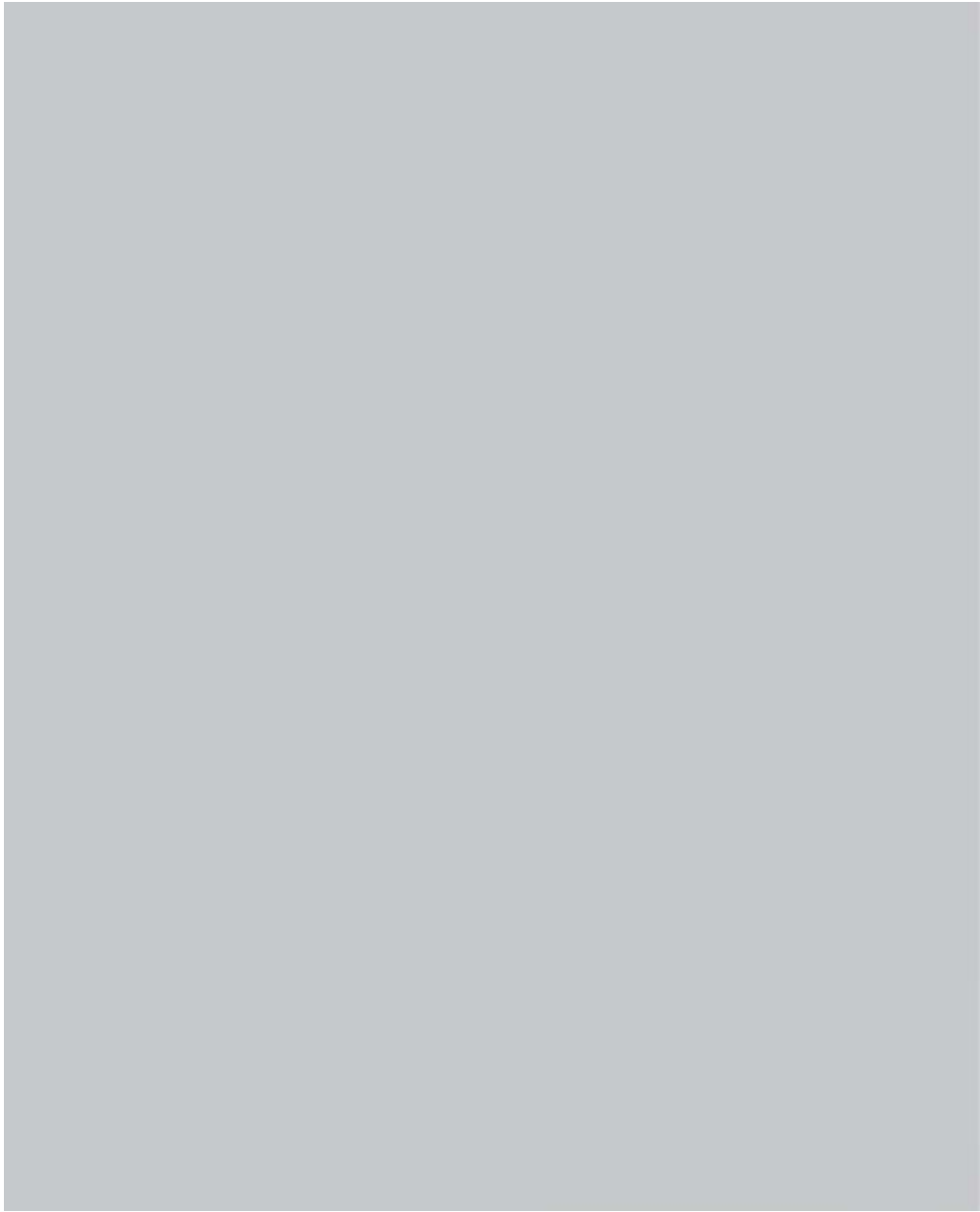
Part 2

Probability Topics

The connection between descriptive and inferential statistics is based on probability concepts. I know the topic of probability theory scares the living daylights out of many students, but it is a very important topic in the world of statistics. The topic of probability acts as a critical link between descriptive and inferential statistics. Without a firm grasp of probability concepts, inferential statistics will seem like a foreign language. Because of this, Part 2 is designed to help you over this hurdle.

I WANT TO LEARN THE PROBABILITY
THEORY, BUT LOOKING AT THE DATA,
I'LL PROBABLY GIVE UP





(hapte

6

Introduction to Probability

In This Chapter

- ◆ Distinguish between classical, empirical, and subjective probability
- ◆ Use frequency distributions to calculate probability <> Examine the basic properties of probability
- ◆ Demonstrate the intersection and union of simple events using a Venn diagram

As we leave the happy world of descriptive statistics, you may feel like you're ready to take on the challenge of inferential statistics. But before we enter that realm, we need to arm ourselves with probability theory. Accurately predicting the probability that an event will occur has widespread applications. For instance, the gaming industry uses probability theory to set odds for lotteries, card games, and sporting events.

The focus of this chapter is to start with the basics of probability, after which we will gently proceed to more complex concepts in Chapters 7 and 8. We'll discuss different types of probabilities and how to calculate the probability of simple events. We'll rely on data from frequency distributions to examine the likelihood of a combination of simple events. So pull up a chair and let's roll those dice!

82 Part 2: Probability Topics

What Is Probability?

Probability concepts surround most of our daily lives. When I see that the weather forecast shows an 80 percent chance of rain tomorrow and I want to play golf or that my beloved Pittsburgh Pirates have only won 40 percent of their games this year (which they also did last year and the year before that), there is a 65 percent chance I will get moody.

In simple terms, probability is the likelihood of a particular event like rain or winning a ballgame. But before we go any further, we need to tackle some new "stat jargon." The following terms are widely used when talking about probability:

Experiment. The process of measuring or observing an activity for the purpose of collecting data. An example is rolling a pair of dice.

Outcome. A particular result of an experiment. An example is rolling a pair of threes with the dice.

Sample space. All the possible outcomes of the experiment. The sample space for our experiment is the numbers {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12}. Statistics people like to put {} around the sample space values because they think it looks cool.

◆ **Event.** One or more outcomes that are of interest for the experiment and which is/are a subset of the sample space. An example is rolling a total of 2, 3, 4, or 5 with two dice.

To properly define probability, we need to consider which type of probability we are referring to.

Classical Probability

Classical probability refers to a situation when we know the number of possible outcomes of the event of interest and can calculate the probability of that event with the following equation:

Number of possible outcomes in which Event A occurs

$P[A] =$

Total number of possible outcomes in the sample space

where:

$P[A]$ = the probability that Event A will occur

Chapter 6: Introduction to Probability 83

For example, if Event A = rolling a total of 2, 3, 4, or 5 with two dice, we need to define the sample space for this experiment, which is shown in the following table.

There are 36 total outcomes for this experiment, each with the same chance of occurring. I have underlined the outcomes that correspond to Event A. There is a total of 10 of them. Therefore:

10

$$P[A] = \frac{10}{36} = 0.28$$

36

definition

Classical probability requires that you know the number of outcomes that pertain to a particular event of interest. You also need to know the total number of possible outcomes in the sample space.

To use classical probability, you need to understand the underlying process so you can determine the number of outcomes associated with the event. You also need to be able to count the total number of possible outcomes in the sample space. As you will see next, this may not always be possible.

Empirical Probability

When we don't know enough about the underlying process to determine the number of outcomes associated with an event, we rely on empirical probability. This type of probability observes the number of occurrences of an event through an experiment and calculates the probability from a relative frequency distribution. Therefore:

Frequency in which Event A occurs

$$P[A] = \frac{\text{Frequency in which Event A occurs}}{\text{Total number of observations}}$$

Total number of observations

Part 2: Probability Topics

definition

•nrnticn

Empirical probability requires that you count the frequency that an event occurs through an experiment and calculate the probability from the relative frequency distribution.

One example of empirical probability is to answer the age-old question "What is the probability that John will get out of bed in the morning for school after his first wake-up call?" Because I cannot begin to understand the underlying process of why a teenager will resist getting out of bed before 2 P.M., I need to rely on empirical probability. The following table indicates the number of wake-up calls John required over the past 20 school days.

John's Wake-Up Calls (Previous 20 School Days)

We can summarize this data with a relative frequency distribution.

Relative Frequency Distribution for John's Wake-Up Calls

Number of Wake-Up Calls

Number of Observations

Percentage

3 4 8 5 Total = 20

$3/20 = 0.15$

$4/20 = 0.20$ $8/20 = 0.40$ $5/20 = 0.25$

Based on these observations, if Event A = John getting out of bed on the first wake-up call, then $P[A] = 0.15$.

Using the previous table, we can also examine the probability of other events. Let's say Event B = John requiring more than 2 wake-up calls to get out of bed; then $P[B] = 0.40 + 0.25 = 0.65$. That boy needs to go to bed earlier on school nights!

Chapter 6: Introduction to Probability 85

Random Thoughts

The probability that you will win a typical state lottery, where you correctly choose 6 out of 50 numbers, is approximately 0.00000006, or 1 out of 16 million. This is calculated using classical probability. Compare this to the probability that you will be struck by lightning once during your lifetime, which is 0.0003 or 1 out of 3,000 (source: www.nws.noaa]. This is an empirical probability determined by the number of times people have been struck by lightning in the past. According to these statistics, you are more than 5,000 times more likely to be struck by lightning than win the lottery! In spite of this, Debbie still makes me go buy a ticket when there's a big jackpot, even during a thunder storm.

If I choose to run another 20-day experiment of John's waking behavior, I would most likely see different results than those in the previous table. However, if I were to observe 100 days of this data, the relative frequencies would approach the true or classical probabilities of the underlying process. This pattern is known as the law of large numbers.

To demonstrate the law of large numbers, let's say I flip a coin three times and each time the result is heads. For this experiment, the empirical probability for the event heads is 100 percent. However, if I were to flip the coin 100 times, I would expect the empirical probability of this experiment to be much closer to the classical probability of 50 percent.

def'i'mi'tiem

The law of large numbers states

that when an experiment is conducted a large number of times, the empirical probabilities of the process will converge to the classical probabilities.

Subjective Probability

We use subjective probability when classical and empirical probabilities are not available. Under these circumstances, we rely on experience and intuition to estimate the probabilities.

Examples where we would apply subjective probability are "What is the probability that my son Brian will ask to borrow my new car, which happens to have a 6-speed manual transmission, for his junior prom?" (97 percent) or "What is the probability that my new car will come back with all 6 gears in proper working order?" (18 percent). I based these probabilities on my personal observations after returning from a

86 Part 2: Probability Topics

"practice run" where I heard noises from my poor transmission that chilled me to the bone and to this day haunt me in my sleep. I need to use subjective probability in this situation because my car would never survive several of these "experiments."

Basic Properties of Probability

Our next step is to review the "rules and regulations" that govern probability theory. The basic ones are as follows:

♦ If $P[A] = 1$, then Event A must occur with certainty. An example is Event A = Debbie buying a pair of shoes this month.

If $P[A] = 0$, then Event A will not occur with certainty. An example is Event A = Bob will eventually finish the basement project that he started three years ago.

♦ The probability of Event A must be between 0 and 1.

♦ The sum of all the probabilities for the events in the sample space must be equal to 1. For example, if the experiment is flipping a coin with Event A = heads and Event B = tails, then A and B represent the entire sample space. We also know that $P[A] + P[B] = 0.5 + 0.5 = 1$.

♦ The complement to Event A is defined as all the outcomes in the sample space that are not part of Event A and is denoted as A' . Using this definition, we can state the following: $P[A] + P[A'] = 1$ or $P[A] = 1 - P[A']$.

For example, if the experiment is rolling a single six-sided die, the sample space is shown in Figure 6.1.

Figure 6.1

Sample space for a single die experiment.

...

Chapter 6: Introduction to Probability 87

If we say that Event A = rolling a 1, then Event A' = rolling a 2, 3, 4, 5, or 6. Therefore:

1

$$P[A] = \frac{1}{6} = 0.167$$

6

$$P[A'] = 1 - 0.167 = 0.833$$

Up to this point, all our examples would be considered cases of simple probability, which is defined as the probability of a single event. Now we'll expand this concept to more than one event.

The Intersection of Events

Sometimes we are interested in the probability of a combination of events rather than just a simple event. To demonstrate this technique, I will use the following example.

Now that my children are older and living away from home, I cherish those moments when the phone rings and I see one of their numbers appear on my caller ID. Experience has taught me that I can categorize these calls as either "crisis," involving such things as a computer, a car, an ATM card, or a cell phone; or "noncrisis," when they call just to see if I'm alive and well enough to help with their next crisis.

The following table, called a contingency table, categorizes the last 50 phone calls by child and type of call.

Contingency Table for Phone Calls

Child Crisis Non-Crisis Totals

Contingency tables show the actual or relative frequency of two types of data at the same time. In this case, the data types are child and type of call.

88 Part 2: Probability Topics

I'll assume that this past pattern of calls will hold true in the near future. We'll define Events A and B as follows:

◆ Event A = the next phone call will come from Christin.

◆ Event B = the next phone call will involve a crisis.

We can use the contingency table to calculate the simple probability that the next phone call will come from Christin as follows:

20

$$P[A] = \frac{20}{50} = 0.40$$

50

The probability that the next phone call will involve a crisis would be as follows:

28

$$P[B] = \frac{28}{50} = 0.56$$

50

defi'iortieio

A contingency table indicates the number of observations that are classified according to two variables. The intersection of Events A and B represents the number of instances where Events A and B occur at the same time (that is, the same phone call is both from Christin and a crisis). The probability of the intersection of two events is known as a joint probability.

What about the probability that the next phone call will come from Christin and will involve a crisis? This event is known as the intersection of Events A and B and is described by $A \cap B$. The number of phone calls from our contingency table that meet both criteria is 14, so:

$$P[A \text{ and } B] = P[A \cap B] = \frac{14}{50} = 0.28$$

This explains why I hold my breath as I pick up the phone!

The probability of the intersection of two events is known as a joint probability.

The Union of Events: A Marriage Made in Heaven

The union of Events A and B represents all the instances where either Event A or Event B or both occur and is denoted as $A \cup B$. Using our previous example, the following table shows the four combinations that include either a call from Christin or a crisis phone call.

Child

Chapter 6: Introduction to Probability 89

Type of Call

Number of Calls

Therefore, the probability that the next phone call is either from Christin or is a crisis is as follows:

$$P[A \text{ and } B] = P[A \cup B] = \frac{34}{50} = 0.68$$

50

defining tier)

The union of Events A and B represents the number of instances where either Event A or B occur (that is, the number of calls that were either from Christin or were a crisis).

Your Turn

1. Define each of the following as classical, empirical, or subjective probability.

- a. The probability that the baseball player Derek Jeter will get a hit during his next at bat.
- b. The probability of drawing an Ace from a deck of cards.

Bob's Basics



The probability of the intersection of two events can

never be more than the probability of the union of two events. If your calculations don't agree with this, go back and check for a mistake!

- c. The probability that I will shoot lower than a 90 during my next round of golf.
- d. The probability of winning the next state lottery drawing.
- e. The probability that the drive belt for my riding lawnmower will break this summer (it did).
- f. The probability that I will finish writing this book before my deadline.

90 Part 2: Probability Topics

2. Identify whether each of the following are valid probabilities.

- a. 65 percent
- b. 1.9
- c. 110 percent
- d. -4.2
- e. 0.75

f. o

3. A survey of 125 families asked whether the household had Internet access. Each family was classified by race. The contingency table is shown here.

A family from the survey is randomly selected. We define:

Event A: The selected family has an Internet connection in its home.

Event B: The selected family is Asian American.

a. Determine the probability that the selected family has an Internet connection.

b. Determine the probability that the selected family is Asian American.

c. Determine the probability that the selected family has an Internet connection and is Asian American.

d. Determine the probability that the selected family has an Internet connection or is Asian American.

4. Using the "crisis" and "noncrisis" phone call example, we define:

Event A: The next phone call will come from Brian.

Event B: The next phone call will be a "noncrisis."

Chapter 6: Introduction to Probability 91

a. Determine the probability that the next phone call will be from Brian and be a "noncrisis."

b. Determine the probability that the next phone call will be from Brian or be a "noncrisis."

The Least You Need to Know

♦ Classical probability requires knowledge of the underlying process in order to count the number of possible outcomes of the event of interest.

- ◆ Empirical probability relies on historical data from a frequency distribution to calculate the likelihood that an event will occur.
- ◆ The law of large numbers states that when an experiment is conducted a large number of times, the empirical probabilities of the process will converge to the classical probabilities.
- ◆ The intersection of Events A and B represents the number of instances where Events A and B occur at the same time.
- ◆ The union of Events A and B represents the number of instances where either Event A or B occur.

Chapter

7

More Probability Stuff

In This Chapter

- ◆ Calculating conditional probabilities
- ◆ The distinction between independent and dependent events
- ◆ Using the multiplication rule of probability
- ◆ Defining mutually exclusive events
- ◆ Using the addition rule of probability
- ◆ Using the Bayes' theorem to calculate conditional probabilities

Now that we have arrived at the second of three basic probability chapters, we're ready for some new challenges. We need to take the probability concepts that you've mastered from Chapter 6 and put them to work on the next step up the ladder. Don't worry if you're afraid of

heights like I am—just keep looking up!

This chapter deals with the topic of manipulating the probability of different events in various ways. As new information about events becomes available, we can revise the old information and make it more useful. This revised information can sometimes lead to surprising results—as you'll soon see.

% Part 2: Probability Topics

Conditional Probability

We define conditional probability as the probability of Event A knowing that Event B has already occurred. To demonstrate this concept, consider this example.

Debbie is an avid tennis player, and we enjoy playing matches against each other. We do, however, have one difference of opinion on the court. Debbie likes to have a nice long warm-up session at the start, where we hit the ball back and forth and back and forth and back and forth. All during this time, a little voice in my head is saying, "Who's winning?" and "What's the score?" My ideal warm-up is to bend at the waist to tie my sneakers and to adjust my shorts. Each tennis match becomes a test of my manhood and the "warm-up" has nothing to do with "the thrill of victory and the agony of defeat." I can't help it; it must be a guy thing that has been passed down through thousands of years of conditioning. Debbie tells me that when we rush through the warm-up, she doesn't play as well. "Poppycock!" I say, and I'll prove it. The following table shows the outcomes of our last 20 matches, along with the type of warm-up before we started keeping score.

Contingency Table for the Tennis Example

Warm-Up Time

Debbie Wins (A)

Bob Wins (A')

Total

Less than 10 min (B) 10 min or more (B') Total

The events of interest are ...

- ◆ Event A = Debbie wins the tennis match.
- ◆ Event B = the warm-up time is less than 10 minutes.
- ◆ Event A = Bob wins the tennis match.
- ◆ Event B' = the warm-up time is 10 minutes or more.

Chapter 7: More Probability Stuff 95

Without any additional information, the simple probability of each of these events is as follows:

$$P[A] = \frac{1}{2} = 0.45$$

$$P[B] = \frac{1}{2} = 0.65$$

1 J 20

$$P[A']$$

$$= 0.35$$

def'i'mi'ticin

11 7

$$\frac{1}{2} = 0.55 \quad P[B'] = \frac{1}{2} =$$

20 L ' 20

As if these probabilities don't have enough names already, I have one

more for you. These are also known as prior probabilities because they are derived only from information that is currently available.

You might wonder, "What other information is he talking about?" Well, suppose I know that we had a warm-up period of less than 10 minutes. Knowing this piece of info, what is the probability that Debbie will win the match? This is the conditional probability of Event A given that Event B has occurred. Looking at the previous table, we can see that Event B has occurred 13 times. Because Debbie has won 4 of those matches (A), the probability of A given B is calculated as follows:

$\frac{4}{13}$ Debbie won't be happy to see that probability.

We can also calculate the probability that Debbie will win, given that the warm-up is 10 minutes or longer (otherwise known as an eternity). According to the previous table, these marathon warm-ups occurred 7 times, with Debbie winning 5 of these matches. Therefore:

Simple or prior probabilities are always based on the total number of observations. In the previous example, it is 20 matches.

$$P[A/B] =$$

$$= 0.31$$

$$P[A/B'] = \frac{5}{7} = 0.71$$

This one looks bad for Bob. I might have to hide this chapter from my live-in proofreader.

96 Part 2: Probability Topics

def

•rri •tier)

Conditional probability is defined as the probability of Event A knowing that Event B has already occurred. Conditional probabilities are also known as posterior probabilities.

Once again, I bring to you more "star, jargon." Conditional probabilities are also known as posterior probabilities (I'll resist using a butt joke here), which are considered revisions of prior probabilities using additional information. For example, the prior probability of Debbie winning is $P[A] = 0.45$. However, with the additional information that the warm-up was 10 minutes or longer, we revise the probability of Debbie winning to $P[A/B'] = 0.71$.

Conditional probabilities are very useful for determining the probabilities of compound events as you will see in the following sections.

Independent Versus Dependent Events

Events A and B are said to be independent of each other if the occurrence of Event B has no effect on the probability of Event A. Using conditional probability, Events A and B are independent of one another if:

$$P[A/B] = P[A]$$

If Events A and B are not independent of one another, then they are said to be dependent events.

In the tennis example, Events A and B are dependent because the probability of Debbie winning depends on whether the warm-up is more or less than 10 minutes. We can also demonstrate this by observing that:

$$P[A] = \frac{9}{20} = 0.45 \text{ and } P[A/B] = \frac{6}{20} = 0.3$$

These probabilities tell us that overall, Debbie wins 45 percent of the matches. However, when there is a short warm-up, she only wins 30 percent of the time. Because these probabilities are not equal, Events A and B are dependent.

An example of 2 independent events is the outcome of rolling two dice:

- ◆ Event A: Roll the number 4 on the first of two dice.
- ◆ Event B: Roll the number 6 on the second of two dice.

For these events, the simple probabilities are as follows:

def # i # ni # tien

$P[A] =$

$: 0.167$

Events A and B are said to be independent of each other if the occurrence of Event B has no effect on the probability of Event A. If Events A and B are not independent of one another, then they are said to be dependent events.

$- = 0.167$ and $P[B] = - = 6/6$

Even if we know that the first die rolled a 4, the probability of the second die being a 6 is not affected because dice, for the most part, are pretty dim-witted and are not very aware of what is going on around them. Knowing this, we can say the following:

$P[B/A] = P[B] = - = 0.167$ 6

Therefore, Events A and B are independent of one another.

Multiplication Rule of Probabilities

We use the multiplication rule of probabilities to calculate the joint probability of two events. In other words, we are calculating the probability of these events occurring at the same time. Chapter 6 referred to this as the intersection of two events. For two independent events, the multiplication rule states the following:

$P[A \text{ and } B] = P[A] \times P[B]$

Recall from Chapter 6 that $P[A \text{ and } B]$ is also known as the joint probability of Events A and B.

For example, we can use the multiplication rule to calculate the joint probability of rolling "snake eyes" with a pair of dice. We define the events as follows:

♦ Event A: Roll a 1 on the first die.

♦ Event B: Roll a 1 on the second die.

Because these events are clearly independent, we can calculate the probability they will occur simultaneously:

$$.6 \text{ J} 6 \text{ J} = 36$$

$$P[A \text{ and } B] =$$

98 Part 2: Probability Topics

def *i •ni , tiein

For dependent events, the multiplication rule states that $P[A \text{ and } B] = P[A/B] \times P[B]$. If the events are independent, the multiplication rule simplifies to $P[A \text{ and } B] = P[A] \times P[B]$.

Bob's Basics



We can rearrange the multiplication rule algebraically and use it to calculate the conditional probability of Event A, given that Event B has occurred, with the following equation:

$$P[A \text{ and } B]$$

$$P[A/B] = -$$

$$P[B]$$

If the two events are dependent, things start to heat up and the multiplication rule becomes:

$$P[A \text{ and } B] = P[A/B] \times P[B]$$

To demonstrate the multiplication rule with dependent events, let's go back to the tennis court and calculate $P[A \text{ and } B]$, the probability that Debbie will win and that the warm-up is less than 10 minutes (from my earlier results):

$$P[B] = 0.65 \text{ and } P[A/B] = 0.31$$

$$P[A \text{ and } B] = (0.65)(0.31)$$

$$P[A \text{ and } B] = 0.20$$

We can confirm this result by checking the original contingency table, where we see that out of 20 matches, Debbie won 4 times with a warm-up of less than 10 minutes. Therefore:

$$P[A \text{ and } B] =$$

$$\frac{4}{20}$$

$$= 0.20$$

Maybe Debbie has a valid complaint after all. I wonder whether she ever gets tired of being right!

Mutually Exclusive Events

Two events are considered to be mutually exclusive if they cannot occur at the same time during the experiment. For example, suppose my experiment is to roll a single die and my events of interest are as follows:

♦ Event A: Roll a 1.

♦ Event B: Roll a 2.

def

•rni'tien

Two events are considered to be mutually exclusive if they cannot occur at the same time during the experiment.

Because there is no way for both of these events to occur simultaneously, they are considered to be mutually exclusive.

Events that can occur at the same time are, you guessed it, not mutually exclusive. In our tennis

Chapter 7: More Probability Stuff 99

example, Events A and B are not mutually exclusive because (a) Debbie can win the match and (b) the warm-up can be less than 10 minutes in the same experiment.

Addition Rule of Probabilities

We use the addition rule of probabilities to calculate the probability of the union of events—that is, the probability that either Event A or Event B will occur. For two events that are mutually exclusive, the addition rule states the following:

def*i # rii # tien

For mutually exclusive events, the addition rule states that $P[A \text{ or } B] = P[A] + P[B]$. If the events are not mutually exclusive, the addition rule becomes $P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$.

$$P[A \text{ or } B] = P[A] + P[B]$$

As an example, for the single-die experiment with mutually exclusive events:

◆ Event A: Roll a 1.

◆ Event B: Roll a 2.

The simple probabilities are as follows:

$$P[A] = \frac{1}{6} = 0.167 \text{ and } P[B] = \frac{1}{6} = 0.167$$

The probability that either a 1 or a 2 will be rolled is as follows:

$$P[A \text{ or } B] = P[A] + P[B] \quad P[A \text{ or } B] = 0.167 + 0.167 \quad P[A \text{ or } B] = 0.334$$

For events that are not mutually exclusive, the addition rule states the following: $P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$

Going back to the tennis court, where ...

◆ Event A = Debbie wins the tennis match.

◆ Event B = The warm-up time is less than 10 minutes.

Recall that:

$$P[A] = 0.45 \text{ and } P[B] = 0.65 \quad P[A \text{ and } B] = 0.20$$

100 Part 2: Probability Topics

Therefore, the probability that Debbie will either win the match or the warm-up will be less than 10 minutes is as follows:

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

$$P[A \text{ or } B] = 0.45 + 0.65 - 0.20$$

$$P[A \text{ or } B] = 0.90$$

The logic behind subtracting $P[A \text{ and } B]$ in the addition rule is to avoid double counting. We can demonstrate this in the following table, which converts the frequency distribution to a relative frequency distribution.

Relative Frequency Distribution for Tennis Matches

Warm-Up Time

Debbie Wins

Bob Wins

Total

Less than 10 10 or more Total

$$4 / 20 = 0.20 \quad 5 / 20 = 0.25 \quad 9 / 20 = 0.45$$

$$9 / 20 = 0.45$$

$$n / w = 0.55$$

$$» / 20 = 0.65$$

$$V_{,,} = 0.35 \quad 20 / 20 = 1.00$$

The union of Events A and B can be displayed using Figure 7.1.

The Union of Events A and B



Bob's Basics

When converting frequencies V_{ij} to relative frequencies in a

contingency table, always divide each number in the table by the total number of observations. In the previous example, that is 20 matches.

The probability of Debbie winning the match (Event A) is represented by the box in the first column. The probability of having a warm-up of less than 10 minutes (Event B) is represented by the box in the first row. If we add $P[A] + P[B]$, which would be the column plus the row in Figure 7.3, we are double counting $P[A \text{ and } B] = 0.20$ and therefore need to subtract this in the addition rule for events that are not mutually exclusive.

Summarizing Our Findings

Before moving on, let's step back and take a look at what we've done so far. Figure 7.2 shows the simple, joint, and conditional probabilities in the relative frequency distribution for our tennis matches.

Summary of Probabilities for Tennis Example

Warm-Up Time

Less than 10 Min (B) 10 Minor More (B')

0.65 0.35

Figure 7.2

Summary of probabilities for the tennis example.

Conditional Probabilities

$P[A \text{ and } B] = 0.20$ $P[A/B] = \frac{0.20}{0.65} = 0.31$

$P[A' \text{ and } B] = 0.45$ $P[A'/B] = \frac{0.45}{0.35} = 0.69$

$P[A \text{ and } B'] = 0.25$

$P[A/B'] = \frac{0.25}{0.65} = 0.71$

$P[B'] = 0.35$

$P[A/B']$

$P[A' \text{ and } B'] = 0.10$ $P[A'/B'] = \frac{0.10}{0.35} = 0.29$

Note that:

- ◆ Event A' = Bob wins the match.
- ◆ Event B' = The warm-up is 10 minutes or more.

These conditional probabilities have revealed my secret to success on the court. The probability of my winning after a short warm-up, $P[A'/B]$, is 0.69; whereas the probability of my winning after a longer warm-up, $P[A'/B']$, is 0.29. I knew I should have picked another example for this chapter.

Bayes' Theorem

Thomas Bayes (1701-1761) developed a mathematical rule that deals with calculating $P[B/A]$ from information about $P[A/B]$. Bayes' theorem states the following:

$$P[B]P[A/B]$$

$$P[B/A] = \text{where:}$$

$$(P[B]P[A/B]) + (P[B']P[A/B'])$$

$P[B']$ = the probability of the complement of Event B

$P[A/B']$ = the probability of Event A, given that the complement to Event B has occurred

Now that looks like a mouthful, but applying it in our tennis example will clear things up. With Bayes' theorem, we can calculate $P[B/A]$, which is the probability that the warm-up was less than 10 minutes, given that Debbie won the match. Using the values from the previous figure:

$$(0.65)(0.31)$$

$$P[B/A] =$$

$$P[B/A] =$$

$$(0.65)(0.31) + (0.35)(0.71)$$

$$0.20$$

$$0.20 + 0.25$$

$$= 0.44$$

Random Thoughts

Not only was Thomas Bayes a prominent mathematician, but he was also a published Presbyterian minister who used mathematics to study religion.

Knowing that Debbie won the match, we can say there is a 44 percent chance that the warm-up was less than 10 minutes.

We can confirm this result by looking at the original contingency table. Because Debbie won 9 matches and from those, 4 had a warm-up of less than 10 minutes:

$P[B/A]$:

0.44

Ta da! Please hold your applause until the end of the book.

Your Turn

A political telephone survey of 260 people asked whether they were in favor or not in favor of a proposed law. Each person was identified as Republican or Democrat. The following contingency table shows the results.

Party	In Favor	Not in Favor	Total
-------	----------	--------------	-------

54	152	
----	-----	--

29	108	
----	-----	--

83	260	
----	-----	--

A person from the survey is selected at random. We define:

◆ Event A: The person selected is in favor of the new law.

◆ Event B: The person selected is a Republican.

1. Determine the probability that the selected person is in favor of the new law.
2. Determine the probability that the selected person is a Republican.
3. Determine the probability that the selected person is not in favor of the new law.

4. Determine the probability that the selected person is a Democrat.
5. Determine the probability that the selected person is in favor of the new law given that the person is a Republican.
6. Determine the probability that the selected person is not in favor of the new law given that the person is a Republican.
7. Determine the probability that the selected person is in favor of the new law given that the person is a Democrat.
8. Determine the probability that the selected person is in favor of the new law and that the person is a Republican.
9. Determine the probability that the selected person is in favor of the new law and that the person is a Democrat.
10. Determine the probability that the selected person is in favor of the new law or that the person is a Republican.

ICi Part 2: Probability Topics

11. Determine the probability that the selected person is in favor of the new law or that the person is a Democrat.
12. Using Bayes' theorem, calculate the probability that the selected person was a Republican, given that the person was in favor of the new law.

The Least You Need to Know

- ◆ We define conditional probability as the probability of Event A knowing that Event B has already occurred.
- ◆ We can say Events A and B are independent of each other if the occurrence of Event B has no effect on the probability of Event A. If Events A and B are not independent of one another, then they are said to be dependent events.
- ◆ For dependent events, the multiplication rule states that

$P[A \text{ and } B] = P[A/B] P[B]$. If the events are independent, the multiplication rule simplifies to $P[A \text{ and } B] = P[A] P[B]$.

♦ We consider two events to be mutually exclusive if they cannot occur at the same time during the experiment.

♦ For mutually exclusive events, the addition rule states that

$P[A \text{ or } B] = P[A] + P[B]$. If the events are not mutually exclusive, the addition rule becomes $P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$.

♦ Bayes' theorem deals with calculating $P[B/A]$ from information about $P[A/B]$ using the following formula:

$$P[B]P[A/B]$$

$$P[B/A] =$$

$$(P[B]P[A/B]) + (P[B']P[A/B'])$$

(hapte

8

Counting Principles and Probability Distributions

In This Chapter

- ♦ Using the fundamental counting principle
- ♦ Distinguishing between permutations and combinations
- ♦ Defining a random variable and probability distribution
- ♦ Calculating the mean and variance of a discrete probability distribution

Well, we've finally arrived at our third and last chapter on general

probability concepts. This chapter sets the stage for the last three chapters in Part 2, which will focus on specific types of probability distributions. Before you know it, we'll be knee deep with inferential statistics.

This chapter will also teach you how to count. This type of counting, however, goes far beyond what you've seen on Sesame Street. Counting events is an important step in calculating probabilities and must be done with care.

106 Part 2: Probability Topics

Counting Principles

To use classical probability, which we introduced way back in Chapter 6, we need to be able to count the number of events of interest along with the total number of events that are possible in the sample space. For simple events, like rolling a single die, the number of possible outcomes (six) is obvious. But for more complex events, like a state lottery drawing, we need to rely on techniques known as counting principles to arrive at the correct answer, so let's look at these techniques.

The Fundamental Counting Principle

After a tough round of golf on a hot afternoon, Brian, John, and I decide to revive our spirits at the ice cream store on the way home. There I'm overwhelmed with deciding

between four flavors and three toppings to indulge in. \f. . . How many different combinations of ice cream and

lACl HOrl , toppings am I faced with? The fundamental counting

principle comes to my rescue by telling me that if one event (my ice cream choice) can occur in m ways and a second event (my topping choice) can occur in n ways, the total number of ways both events can occur together is $m \times n$ ways. In my case, I have $m \times n$ combinations of flavors and toppings in which to blow my diet. (I'll leave that topic for another chapter.)

According to the fundamental counting principle, if one event can occur in m ways and a second event can occur in n ways, the total number of ways both events can occur together is $m \cdot n$ ways. And we can extend this principle to more than two events.

Now I can extend this principle to more than two events. In addition to flavors and toppings, I have another tempting choice between a small and large serving. That leaves me with the mind-boggling decision of $4 \cdot 3 \cdot 2 = 24$ combinations, which are summarized in the table that follows my list of options.

List of Combinations (Flavor-Topping-Size)

CH-HF-LG VA-HF-LG ST-HF-LG CF-HF-LG

CH-HF-SM VA-HF-SM ST-HF-SM CF-HF-SM

CH-BS-LG VA-BS-LG ST-BS-LG CF-BS-LG

CH-BS-SM VA-BS-SM ST-BS-SM CF-BS-SM

CH-SP-LG VA-SP-LG ST-SP-LG CF-SP-LG

CH-SP-SM VA-SP-SM ST-SP-SM CF-SP-SM

Can you guess which choice a certain chocolate-loving author made?

Another demonstration of the fundamental counting principle is to calculate the number of unique combinations for a state's automobile license plates. Suppose the state plates have three letters followed by four numbers. The number zero and the letter O are not eligible because their resemblance may cause confusion. Because we have 25 possible letters and 9 possible numbers, the total number of unique combinations is as follows:

$25 \times 25 \times 25 \times 9 \times 9 \times 9 \times 9 = 102,515,625$! That's 102,515,625 possible combinations!

Permutations

Permutations are the number of different ways in which objects can be arranged in order. In a permutation, each item appears only once. The number of permutations of n distinct objects is $n!$ (expressed as n factorial) and is defined as follows:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots 4 \cdot 3 \cdot 2 \cdot 1$$

108 Part 2: Probability Topics

By definition, $0! = 1$. For instance, $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$. As an example, there are six permutations for the numbers 1, 2, and 3, as shown here:

123 132 213 231 312 321

Because: $3! = 3 \cdot 2 \cdot 1 = 6$

definition

Permutations are the number of different ways in which objects can be arranged in order. The number of permutations of n objects taken r at a time can

be found by $nPr = \frac{n!}{(n-r)!}$.

$n! = n \cdot (n-1) \cdot (n-2) \cdots 1$

Before the beginning of a professional basketball game, the starting 5 players are announced one at a time. How many different ways can we arrange the order that the players are announced? The number of permutations is:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

Suppose we want to select only some of the objects in the group. The number of permutations of n objects taken r at a time can be found as follows:

$$(n - r)! = (n - r) \cdot (n - r - 1) \cdots 1$$

Bob's Basics



It's easier to calculate the number of permutations using this formula:

$${}_nP_r = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)$$

This works because every value in the denominator (the bottom of the fraction) will

cancel out with many values in the numerator (the top of the fraction).

Using our basketball example again, if there are 12 players on the team, how many different ways can any five players on the team be announced to start the game? In this case, because $n = 12$ and $r = 5$, the number of permutations is as follows:

$$12! / (12-5)! = 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 / 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

${}_nP_r =$

$${}_{12}P_5$$

$$(12-5)! \cdot 12!$$

$$7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

$$12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 = 95,040$$

$(12-5)!$ I'm sure glad it's not my job to decide who gets announced first.

Sometimes the order of events is not of consequence, and we'll discuss those cases in the next section.

Combinations

Combinations are similar to permutations, except that the order of the objects is not important. The number of combinations of n objects taken r at a time can be found as follows:

$${}_nC_r$$

$(n-r)!/r!$

For example, in poker, five cards are selected randomly from a deck of 52 cards. How many five-card combinations exist?

$${}^nC_5 = \frac{52!}{5! \cdot 47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2,598,960$$

How many five-card permutations exist?

$52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$

definition

Combinations are the number of different ways in which objects can be arranged without regard to order. The number of combinations of n objects taken r at a time can be found

by nC_r ,

$(n-r)!/r!$

Example ■■

$(52-5)!$

$$= \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 311,875,200$$



n

Bob's Basics

It's easier to calculate the number of combinations using the same logic mutation formula and this formula:

$$n! / (n-r)! = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)$$

per-

${}^nC_r =$

$$(n-r)!r!$$

There are more five-card permutations because the following two poker hands would be considered two different permutations but be counted as only one combination because they are the same cards only in different order.

110 Part 2: Probability Topics

Hand 1

Hand 2

Ace of Spades Queen of Hearts Ten of Spades Ten of Diamonds Three of Clubs

Ace of Spades Ten of Spades Queen of Hearts Ten of Diamonds Three of Clubs

Now that we know the total number of five-card combinations from a 52-card deck, we can calculate the probability of a flush, which is any five cards that are all the same suit (spades, clubs, hearts, or diamonds). For you poker veterans, I am including a royal flush and a straight flush in this calculation. First, we need to count the number of five-card flushes of one suit, let's say diamonds. Because there are 13 diamonds in the deck, the number of combinations of these 13 diamonds, taken five at a time, is as follows:

$${}_{13}P_5 = \frac{13!}{13-5} = \frac{13!}{8!}$$

$${}_{13}C_5$$

$$\frac{(13-5)!5!}{5!5!}$$

$$5-4-3-2-1$$

$$= 1,287$$

Because there are four suits in the deck, the total number of five-card flushes from any suit is $1287 \times 4 = 5,148$. Therefore, the probability of

being dealt a flush, including royal and straight, in a five-card hand is:
5,148

$P[\text{Flush}] =$

2,598,960

$= 0.002$

or roughly twice in 1,000 hands of poker. Ready to deal?

Bob's Basics



An alternate notation for „C,

which you may

come across in other textbooks. Statisticians just love to have different notations for the same concept!

What about the probability of being dealt a hand with two pairs of any suit? There are $13C2 = 78$ different two-pair combinations in the deck. Each pair can have $4C2 = 6$ different combinations of the four suits. There are $52 - 6 = 44$ possible cards left for the fifth card in the hand. The number of two-pair hands would then be:

$78 \cdot 6 \cdot 6 \cdot 44 = 123,552$

Therefore, the probability of being dealt two pair is

$P[\text{Two-pair}] = \frac{123,552}{2,598,960} = 0.0475$

Combinations are also useful for calculating the probability of winning a state lottery

drawing. A typical lottery game requires you to pick six numbers out of a possible 49.

Because the order of the numbers does not matter, we use the combination rather

than the permutation formula. The number of six-number combinations from a pool

of 49 numbers is this:

49C6 = 49! / (6! * 43!) = 13,983,816

$$\frac{49!}{6! \cdot 43!} = 13,983,816$$

$$\frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13,983,816$$

Because there are nearly 14 million different six-number combinations, the probability that your combination is the winner is as follows:

$$P[\text{Winning a 6/49 Lottery}] = \frac{1}{13,983,816} \approx 0.00000007$$

$$\frac{1}{13,983,816}$$

*)

\y^, Wrong Number

^- / ^P Probability does not have a memory. The same six numbers selected in last

week's lottery drawing have the exact same probability of being chosen again in this week's lottery. That's because the two drawings are independent events and have absolutely no influence on each other. Therefore, choosing a lottery number because it has not been selected recently does not increase your odds of winning. Sorry if I ruined your favorite strategy!

With those chances of winning the lottery, you better not quit your day job just yet.

Using Excel to Calculate Permutations and Combinations

Here's something that's pretty cool—rather than deal with all those nasty

factorial calculations, we can let Excel figure out the number of permutations or combinations for us. The functions are:

=PERMUT(z, r)

=COMBIN(z, r)

For example, if we type =PERMUT(12,5) into Excel, the result will be 95,040. Be sure to try this and give your poor calculator a rest!

112 Part 2: Probability Topics

Random Thoughts

The odds of winning the state lottery drawing are so astronomically low, it's hard to really fathom them. Using the 6/49 lottery example, if I bought one ticket every day of the year, I can expect to win once every 38,312 years. To give this some perspective, 38,000 years ago, people were living in caves during the Stone Age. I'm not sure I want to wait that long, no matter how much money I win.

As we wrap up the topic of counting principles, many of you may be surprised at how complicated it can be to count events. But this is an important concept in statistics that we will revisit in Chapter 9.

Probability Distributions

Now let's introduce you to probability distributions and prepare you for the last three chapters of Part 2. However, first we need to discuss the topic of random variables, which will lay the groundwork for specific probability distributions in Chapters 9, 10, and 11.

def

Random Variables

In Chapter 6, we talked about conducting experiments to acquire data. Examples of experiments could be rolling dice or counting the number of times next month that I can't find something in the house and need to ask Debbie to help me. "She who knows where all things are" has this

mystical ability to make these items suddenly appear before my very eyes after I have given up looking. Debbie then proceeds to give me a pitiful look that says, "You would never survive a single day in this world without me," which sadly I'd have to agree with.

The outcomes of these experiments are considered random variables. By definition, these outcomes are not known before the experiment. For example, I can't predict with certainty the number of times next month I'll need Debbie's help finding something. Once the outcome has occurred, I can determine the value of the random variable. For instance, if I ask Debbie to help me four times next month, the value of that random variable is four.

•i*mi*ticn

A random variable is an outcome that takes on a numerical value as a result of an experiment. The value of the random variable, which is not known with certainty before the experiment, is often denoted by x .

All random variables are not created equal. The first type are known as continuous random variables, which are the result of a measurement on a continuous number scale. For example, each morning when I take a deep breath and step on the bathroom scale to weigh myself (taking a deep breath and holding it somehow makes me feel lighter), I'm looking down in shock and disbelief at a continuous random variable. (Maybe I should have chosen the small Death by Chocolate serving.) Examples of values for continuous random variables of this sort could be 180, 181.5, 183.2, and so on. (I'll stop there.) Because this is a continuous variable, my morning weight could take on an unlimited number of possible values, which is a very disconcerting thought.

The second type of random variable is discrete. Discrete random variables are the

result of counting outcomes rather than measuring them. Discrete random variables

can only take on a certain number of integer

values within an interval. An example of a

discrete random variable would be my golf score for my next round because this value is arrived at by counting my total strokes over 18 holes of play. Obviously, this value needs to be an integer, such as 94, because there is no way to count a partial stroke (even though there are times my golf swing feels like one).

def'i'mi'ticm

We will discuss continuous random variables in more detail in Chapter 11. But here and in Chapters 9 and 10 we will focus solely on discrete random variables.

A random variable is continuous if it can assume any numerical value within an interval as a result of measuring the outcome of an experiment. A random variable is discrete if it is limited to assuming only specific integer values as a result of counting the outcome of an experiment.

Discrete Probability Distributions

A listing of all the possible outcomes of an experiment for a discrete random variable along with the relative frequency or probability of each outcome is called a discrete probability distribution. To illustrate this concept, I'll use this example.

My oldest daughter, Christin, was a very accomplished competitive swimmer between the ages of 7 and 13, but her talent certainly didn't come from my side of the family. One day, I mustered the courage to ask Christin to teach me how to swim the butterfly stroke. My form was best described as "a beached whale having seizures." The lifeguards banned me from ever attempting this stroke again, claiming it too closely

resembled a person who was drowning. Somehow, in spite of this gene pool (Get it? Swimming pool, gene pool?), Christin could not only swim, but she also could swim fast.

IK Part 2: Probability Topics

The following table is a relative frequency distribution showing the number of first-second-, third-, fourth-, and fifth-place finishes Christin earned during 50 races.

Place

Number of Races Relative Frequency (Probability)

27

12

7

3

1

Total = 50

$$27 / 50 = 0.54 \%, = 0.24 \text{ y } 50 = 0.14 \text{ V so } = 0.06 \text{ V } 50 = 0.02 \text{ Total} = 1.00$$

If we define the random variable x = the place Christin finished in a race, the previous table would be the discrete probability distribution for the variable x . From this table, we can state the probability that Christin will finish first as follows:

$$P[* = 1] = 0.54$$

Or we can state the probability that Christin will finish either first or second as follows:

$P[x = 1 \text{ or } x = 2] = 0.54 + 0.24 = 0.78$ Figure 8.1 shows the discrete probability distribution for x graphically.

0.6 0.5

Figure 8.1

The discrete probability distribution for Christin's races.

0.4

0.3

0.2

0.1

0

First

Second

Third

Rules for Discrete Probability Distributions

Any discrete probability distribution needs to meet the following requirements:

◆ Each outcome in the distribution needs to be mutually exclusive—that is, the value of the random variable cannot fall into more than one of the frequency distribution classes. For example, it is not possible for Christin to take first and second place in the same race.

◆ The probability of each outcome, $P[r]$, must be between 0 and 1; that is,

$0 < P[r] < 1$ for all values of r . In the previous example, $P[x = 3] = 0.14$, which falls between 0 and 1.

◆ The sum of the probabilities for all the outcomes in the distribution needs to add up to 1; that is, $\sum P[x] = 1$. In the swimming example, the sum of the Relative Frequency (Probability) column in the previous table adds up to 1.

The Mean of a Discrete Probability Distribution

The mean of a discrete probability distribution is simply a weighted average (discussed in Chapter 4) calculated using the following formula:

where:

μ = the mean of the discrete probability distribution
 x_i = the value of the random variable for the i 'th outcome
 $P[x_i]$ = the probability that the i th outcome will occur
 n = the number of outcomes in the distribution

The table that follows revisits Christin's swimming probability distribution. Place x in the first column and $P[x]$ in the second column.

1 0.54

2 0.24

3 0.14

4 0.06

5 0.02

116 Part 2: Probability Topics

The mean of this discrete probability distribution is as follows:

$$\mu = (1)(0.54) + (2)(0.24) + (3)(0.14) + (4)(0.06) + (5)(0.02)$$

$$\mu = 1.78$$

This mean is telling us that Christin's average finish for a race is 1.78 place! How does she do that? Obviously, this will never be the result of any one particular race. Rather, it represents the average finish of many races. The mean of a discrete probability distribution does not have to equal one of the values of the random variable (1, 2, 3, 4, or 5 in this case).

Another term for describing the mean of a probability distribution is the expected value, $E[x]$. Therefore:

$$E[x] = \sum_{i=1}^n x_i V[x_i]$$

An expected value is the mean of a probability distribution.

Didn't say statisticians love all sorts of notation to describe the same concept?

The Variance and Standard Deviation of a Discrete Probability Distribution

Just when you thought it was safe to get back into the water, along comes another variance! Well, if you've seen one variance calculation, you've seen them all. You can calculate the variance for a discrete probability distribution as follows:

where:

σ^2 = the variance of the discrete probability distribution

As before, the standard deviation of the distribution is as follows:

To demonstrate the use of these equations, we'll rely on Christin's swimming distribution. The calculations are summarized in the following table.

$$\sigma^2 = \sum_{i=1}^n (x_i - \mu)^2 P[x_i] = 1.052$$

$$\sigma = 1$$

The standard deviation of this distribution is: $\sigma = \sqrt{1.052} = 1.026$

A more efficient way to calculate the variance of a discrete probability distribution is:

$$\sigma^2 = \sum_{i=1}^n x_i^2 P[x_i] - \mu^2$$

The following table summarizes these calculations using Christin's swimming example.

$x, P(x), \dots, P(x) / x^t$

1 0.54 1 0.54

2 0.24 4 0.96

3 0.14 9 1.26

4 0.06 16 0.96

5 0.02 25 0.50

$\sigma^2 = P(x) = 4.22 - \frac{1}{n} \sum x^2 = 4.22 - \frac{1}{5} (0.54 + 0.96 + 1.26 + 0.96 + 0.50) = 1.052$

$\sigma^2 = 4.22 - (1.78)^2 = 1.052$ As you can see, the result is the same, but with less effort!

118 Part 2: Probability Topics

Your Turn

1. A restaurant has a menu with three appetizers, eight entrees, four desserts, and three drinks. How many different meals can you order?
2. A multiple-choice test has 10 questions, with each question having four choices. What is the probability that a student, who randomly answers each question, will answer each question correctly?
3. The NBA teams with the 13 worst records at the end of the season participate in a lottery to determine the order in which they will draft new players for the next season. How many different arrangements exist for the drafting order for these 13 teams?
4. In a race with eight swimmers, how many ways can swimmers finish first, second, and third?
5. How many different ways can 10 new movies be ranked first and second by a movie critic?
6. A combination lock has a total of 40 numbers and will unlock with the proper three-number sequence. How many possible combinations exist?

7. I would like to select three paperback books from a list of 12 books to take on vacation. How many different sets of three books can I choose?

8. A panel of 12 jurors needs to be selected from a group of 50 people. How many different juries can be selected?

9. A survey of 450 families was conducted to find how many cats were owned by each respondent. The following table summarizes the results.

Number of Cats	Number of Families
----------------	--------------------

0	137
---	-----

1	160
---	-----

2	112
---	-----

3	31
---	----

4	10
---	----

Develop a probability distribution for this data and calculate the mean, variance, and standard deviation.

10. What is the probability of being dealt a full house (three-of-a-kind and a pair) in five-card poker?

The Least You Need to Know

◆ The fundamental counting principle states that if one event can occur in m ways and a second event can occur in n ways, the total number of ways both events can occur together is $m \cdot n$ ways. We can extend this principle to more than two events.

◆ Permutations are the number of different ways in which objects can be arranged in order. Combinations are the number of different ways in which objects can be arranged when order is of no importance.

◆ A probability distribution is a listing of all the possible outcomes of an experiment along with the relative frequency or probability of each

outcome.

♦ A random variable is an outcome that takes on a numerical value as a result of an experiment. The value is not known with certainty before the experiment.

♦ A random variable is continuous if it can assume any numerical value within an interval as a result of measuring the outcome of an experiment. A random variable is discrete if it is limited to assuming only specific integer values as a result of counting the outcome of an experiment.

♦ You find the mean of a discrete probability distribution as follows:

$$\bar{x} = \sum x_i P_i$$

♦ You find the variance of a discrete probability distribution as follows:

$$\sigma^2 = \sum x_i^2 P_i - (\bar{x})^2$$

Chapter

The Binomial Probability Distribution



In This Chapter

- ♦ Describe the characteristics of a binomial experiment
- ♦ Calculate the probabilities for a binomial distribution
- ♦ Find probabilities using a binomial table
- ♦ Find binomial probabilities using Excel
- ♦ Calculate the mean and standard deviation of a binomial distribution

Our discussion of discrete probability distributions so far has been limited to general distributions based on historical data that has been previously collected. However, some theoretical probability distributions are based on a mathematical formula rather than historical data. We will address the first of these, the binomial probability distribution, in this chapter.

In many types of problems we are interested in the probability of an event occurring several times. A classical example that has been torturing students for many years is "What is the chance of getting 7 heads when tossing a coin 10 times?" By the time you finish this chapter, answering this question will be a piece of cake!

122 Part 2: Probability Topics

Characteristics of a Binomial Experiment

If you remember, in Chapter 6 we defined experimenting as the process of measuring or observing an activity for the purpose of collecting data. Let's say our experiment of interest involves a certain professional basketball player shooting free throws. Each free throw would be considered a trial for the experiment. For this particular experiment, we have only two possible outcomes for each trial; either the free throw goes in the basket (a success) or it doesn't (a failure). Because we can have only two possible outcomes for each trial, this is known as a binomial experiment.

Let's say that our player of interest is Michael Jordan, who historically has made 80 percent of his free throws. So the probability of success, p , of any given free throw is 0.80. Because there are only two outcomes possible, the probability of failure for any given free throw, q , is 0.20. For a binomial experiment, the values of p and q must be the same for every trial in the experiment. Because only two outcomes are allowed in a binomial experiment, $p = 1 - q$ always holds true.

def # i*ni # tien

A binomial experiment has the following characteristics: (1) the experiment consists of a fixed number of trials denoted by n ; (2) each trial has only two possible outcomes, a success or a failure; (3) the

probability of success and the probability of failure are constant throughout the experiment; (4] each trial is independent of any other trial in the experiment.

Random Thoughts

Binomial experiments are also known as Bernoulli process, named after Swiss mathematician James Bernoulli, who lived during the 1600s. Repeating a Bernoulli process several times is referred to as Bernoulli trials, a concept that has been haunting students for hundreds of years!

Finally, a binomial experiment requires that each trial is independent of any other trial. In other words, the probability of the second free throw being successful is not affected by whether the first free throw was successful. Other examples of binomial experiments include the following:

- ◆ Testing whether a part is defective after it has been manufactured
- ◆ Observing the number of correct responses in a multiple-choice exam
- ◆ Counting the number of American households that have an Internet connection

Now that we have defined the ground rules for binomial experiments, we are ready to graduate to calculating binomial probabilities.

The Binomial Probability Distribution

The binomial probability distribution allows us to calculate the probability of a specific number of successes for a certain number of trials. Therefore, the random variable for this distribution would be the number of successes that were observed. To demonstrate a binomial distribution, I will use the following example.

Debbie has trained our dog, Kaylee, to do an incredible trick. First thing every morning after she lets the dog out the back door, Kaylee runs like greased lightning around the house, down our rather long driveway, grabs our newspaper, and races to the back door, where she dutifully deposits it on the step. In return for this vital chore for our household,

she gets two cups of dry dog food in a plastic bowl. Amazing, you say. But you've only heard half of it. Somehow, in the tiny recesses of Kaylee's doggy brain, she has worked out the remarkable deduction that "two breakfasts are better than one" and at every opportunity goes on a neighborhood hunt for more newspapers to deposit on our back step. Once she dragged an entire phone book back, thinking maybe this would earn her a bonus. We have failed miserably trying to train Kaylee to return these papers—apparently tiny doggy brains don't work in reverse.

So my job on many afternoons is to carefully

return the stolen merchandise, hoping Kij

$[n - r]!$

represents the number of combinations of n objects taken r at a time.

my neighbors fail to notice the dog slobber // t\ Remember from Chapter 8

on their three-day-old paper. Anyway, let's If a $n \sim$

■ $i \rightarrow n$ that $r \sim r - \gg Ct$, which

say on any particular day there is a 30 percent probability that Kaylee will bring back one stolen paper and a 70 percent chance that she won't. We will assume that she will not bring back more than one paper a day. This scenario represents a binominal experiment, with each day being a Bernoulli trial with $p = 0.30$ (the probability of a "success") and $q = 0.70$ (the probability of a "failure"). We can calculate the probability of r successes in n trials using the binomial distribution, as follows:

$PM] = 7 \cdot P(r)$

$[n - r]!$

With this equation, we can calculate the probability that Kaylee will bring back three papers over the next five days.

$P[3,5] = 5! (0.3)^3 (0.7)^{5-3} / 3! \cdot 2!$

12<t Part 2: Probability Topics

$$P[3,5] = (- j(0.027)(0.49) = 0.1323$$

There is a 13 percent chance that the neighborhood paper bandit will strike 3 times during the next 5 days. We can also calculate the probability that she will round up zero, one, two, four, or five papers over the next five days.

For $r = 0$:

$$P[0,5] =$$

$$P[0,5] = \text{For } r = 1: P[1,5] =$$

$$\frac{5!}{(5-r)!r!} (0.027)^r (0.49)^{5-r}$$

$$(5-0)!0!$$

$$= \frac{5!}{5!0!} (0.027)^0 (0.49)^5$$

$$= \frac{120}{120} (0.027)^0 (0.49)^5$$

$$= 0.00012$$

$$;0.3) u (0.7 r$$

$$;i)(0.1681) = 0.1681$$

$$(0.3)^1(0.7)^4 = 0.2401$$

$$1, |I.?j:=|^{\wedge}|(0.3)(0.2401) = 0.3601$$

For $r = 2$:

$$P[2,5] =$$

$$= \frac{5!}{(5-2)!2!} (0.027)^2 (0.49)^3$$

$$= \frac{120}{2!120} (0.027)^2 (0.49)^3$$

$$; 0.3)^2 (0.7)^3$$

$$P[2,5] = \binom{5}{2} (0.3)^2 (0.7)^3 = 0.3087$$

For $r = 4$:

$$P[4,5] =$$

$$\frac{(5-4)!4!}{5!}$$

$$= \frac{1 \cdot 24}{120}$$

$$; 0.3)^4 (0.7)^1 = 0.0081$$

$$P[4,5] = \binom{5}{4} (0.3)^4 (0.7)^1 = 0.0283$$

For $r = 5$:

$$P[5,5] =$$

$$\frac{5!}{5!}$$

$$= \frac{1}{1} = 0.0024$$

$$P[5,5] = \binom{5}{5} (0.3)^5 (0.7)^0 = 0.0024$$

The following table summarizes all the previous probabilities.

$$P[r,5]$$

$$0.1681 \quad 0.3601 \quad 0.3087 \quad 0.1323 \quad 0.0283 \quad 0.0024 \quad \text{Total} = 1.0$$

This table represents the binomial probability distribution for r successes in five trials with the probability of success equal to 0.30. Notice that the sum of all the probabilities equals 1, which is a requirement for all probability distributions. Figure 9.1 shows this probability distribution as a histogram.

Figure 9.1

distribution.

1 2 3

Number of Successes

126 Part 2: Probability Topics

From this figure, we can see that the most likely number of papers that Kaylee will show up with over 5 days is 1.

Finally, we can calculate the probability of multiple events for this distribution. For instance, the probability that Kaylee will steal at least three papers over the next five days is this:

$$P[r > 3] = P[3,5] + P[4,5] + P[5,5]$$

$$P[r > 3] = 0.1323 + 0.0283 + 0.0024 = 0.163$$

Also, the probability that Kaylee will take no more than one paper over the next five days is this:

$$P[r \leq 1] = P[0,5] + P[1,5]$$

$$P[r \leq 1] = 0.1681 + 0.3601 = 0.5285$$

Our neighbors will be thrilled to see these figures!

Binomial Probability Tables

As the number of trials increases in a binomial experiment, calculating probabilities using the previous formula will really drain the batteries in your calculator and possibly even your brain. An easier way to arrive at these probabilities is to use a binomial probability table, which I have conveniently provided in Appendix B of this book. Below is an excerpt from this appendix, with the probabilities from our previous example underlined.

The probability table is organized by values of n , the total number of trials. The number of successes, r , are the rows of each section, whereas the probability of success, p , are the columns. Notice that the sum of each block of probabilities for a particular value of p adds to 1.0.

One limitation of using binomial tables is that you are restricted to using only the values of p that are shown in the table. For instance, the previous table would not be useful for $p = 0.35$. Other statistics books may contain binomial tables that are more extensive than the one in Appendix B.

Using Excel to Calculate Binomial Probabilities

A convenient way to calculate binomial probabilities is to rely on our friend Excel, with its BINOMDIST function. This built-in function has the following characteristics:

`BINOMDIST(r, n, p, cumulative)`

where:

`cumulative = FALSE` if you want the probability of exactly r successes

`cumulative = TRUE` if you want the probability of r or fewer successes

For instance, Figure 9.2 shows the BINOMDIST function being used to calculate the probability that Kaylee will bring back exactly two papers during the next five days.

128 Part 2: Probability Topics

Figure 9.2

BINOMDIST function in Excel for exactly r successes.

Cell A1 contains the Excel formula `=BINOMDIST(2,5,0.3,FALSE)` with the result being 0.3087.

Excel will also calculate the probability that Kaylee will bring back no more than two papers over the next five days, as shown in Figure 9.3.

Figure 9.3

BINOMDIST function in Excel for no more than r successes.

Cell A1 contains the Excel formula `=BINOMDIST(2,5,0.3,TRUE)` with the result being 0.8369, which is the same as this:

$$P[r < 2] = P[0,5] + P[1,5] + P[2,5]$$

$$P[r < 2] = 0.1681 + 0.3601 + 0.3087 = 0.8369$$

In other words, there is more than an 83 percent chance Kaylee will show up at our back door with 0, 1, or 2 papers that don't belong to us during the next 5 days. That dog sure does keep me busy!

One benefit of using Excel to determine binomial probabilities is that you are not limited to the values of p shown in the binomial table in Appendix B. Excel's BINOMDIST function allows you to use any value between 0 and 1 for p .

The Mean and Standard Deviation for the Binomial Distribution

You can calculate the mean for a binomial probability distribution by using the following equation:

where:

n = the number of trials

p = the probability of a success

For Kaylee's example, the mean of the distribution is as follows: $\mu = np = (5)(0.3) = 1.5$ papers

In other words, Kaylee brings back, on average, 1.5 papers every 5 days.

You can calculate the standard deviation for a binomial probability distribution using the following equation:

$$\sigma = \sqrt{npq}$$

where:

q = the probability of a failure

For our example, the standard deviation of the distribution is as follows:
 $\sigma = \sqrt{npq} = \sqrt{(5)(0.3)(0.7)} = 1.02$ papers

Well, that about covers the binomial probability distribution discussion. Don't be too sad, though; you'll see this again in future chapters.

Your Turn

1. What is the probability of seeing exactly 7 heads after tossing a coin 10 times?
2. Goldey-Beacom College accepts 75 percent of applications that are submitted for entrance. What is the probability that they will accept exactly three of the next six applications?

130 Part 2: Probability Topics

3. Michael Jordan makes 80 percent of his free throws. What is the probability that he will make at least six of his next eight free-throw attempts?
4. A student randomly guesses at a 12-question, multiple-choice test where each question has 5 choices. What is the probability that the student will correctly answer exactly six questions?
5. Historical records show that 5 percent of people who visit a particular website purchase something. What is the probability that no more than two people out of the next seven will purchase something?
6. During the 2005 Major League Baseball season, Derrek Lee had a 0.335 batting average. Construct a binomial probability distribution for the number of successes (hits) for four official at bats during this season.
7. Sixty percent of a particular college population are female students. What is the probability that a class of 10 students has exactly 4 female students?

The Least You Need to Know

- ◆ A binomial experiment has only two possible outcomes for each trial.
- ◆ For a binomial experiment, the probability of success and failure is constant.

Each trial of a binomial experiment is independent of any other trial in the experiment.

♦ The probability of r successes in n trials using the binomial distribution is as follows:

n

$$P(r) = \binom{n}{r} p^r q^{n-r}$$

$(n-r)!$

♦ Calculate the mean for a binomial probability distribution by using the equation

$$\mu = np$$

♦ Calculate the standard deviation for a binomial probability distribution by using the equation $\sigma = \sqrt{npq}$.

(Chapter



The Poisson Probability Distribution

In This Chapter

- ♦ Describe the characteristics of a Poisson process
- ♦ Calculate probabilities using the Poisson equation
- ♦ Use the Poisson probability tables
- ♦ Use Excel to calculate Poisson probabilities
- ♦ Use the Poisson equation to approximate the binomial equation

Now that we have mastered the binomial probability distribution, we are ready to move on to the next discrete theoretical distribution, the Poisson. This probability distribution is named after Simeon Poisson, a French mathematician who developed the distribution during the early 1800s.

The Poisson distribution is useful for calculating the probability that a certain number of events will occur over a specific period of time. We could use this distribution to determine the likelihood that 10 customers will walk into a store during the next hour or that 2 car accidents will occur at a busy intersection this month. So let's grab some crepes and croissants and learn about some French math.

132 Part 2: Probability Topics

Characteristics of a Poisson Process

In Chapter 9, we defined a binomial experiment, otherwise known as a Bernoulli process, as counting the number of successes over a specific number of trials. The result of each trial is either a success or a failure. A Poisson process counts the number of occurrences of an event over a period of time, area, distance, or any other type of measurement.

Rather than being limited to only two outcomes, the Poisson process can have any number of outcomes over the unit of measurement. For instance, the number of customers who walk into our local convenience store during the next hour could be zero, one, two, three, or so on. The random variable for the Poisson distribution would be the actual number of occurrences—in this case, the number of customers arriving during the next hour.

The mean for a Poisson distribution is the average number of occurrences that would be expected over the unit of measurement. For a Poisson process, the mean has to be the same for each interval of measurement. For instance, if the average number of customers walking into the store each hour is 11, this average needs to apply to every one-hour increment.

The last characteristic of a Poisson process is that the number of occurrences during one interval is independent of the number of occurrences in other intervals. In other words, if six customers walk into

the store during the first hour of business, this would have no effect on the number of customers arriving during the second hour.

def'i'ni'tiəm

A Poisson process has the following characteristics: (1) the experiment consists of counting the number of occurrences of an event over a period of time, area, distance, or any other type of measurement; (2) the mean of the Poisson distribution has to be the same for each interval of measurement; (3) the number of occurrences during one interval is independent of the number of occurrences in any other interval.

Examples of random variables that may follow a Poisson probability distribution include the following:

- ◆ The number of cars that arrive at a tollbooth over a specific period of time
- ◆ The number of typographical errors found in a manuscript
- ◆ The number of students who are absent in my Monday-morning statistics class
- ◆ The number of professional football players who are placed on the injured list each week

Now that you understand the basics of a Poisson process, let's move into probability calculations.

The Poisson Probability Distribution

If a random variable follows a pattern consistent with a Poisson probability distribution, we can calculate the probability of a certain number of occurrences over a given interval. To make this calculation, we need to know the average number of occurrences for the event over this interval. To demonstrate the use of the Poisson probability distribution, I'll use this example.

The following story is true, but the names have not been changed because nobody in this story is innocent. Not that any of the previous stories have

been false, but this one is "really" true. Each year, Brian, John, and I make a golf pilgrimage to Myrtle Beach, South Carolina. On our last night one particular year, we were browsing through a golf store. Brian somehow convinced me to purchase a used, fancy, brand-name golf club that he swore he absolutely had to have in order to reach his full potential as a golfer. Even used, this club cost more than any I had purchased new, but teenagers have this special talent that allows them to disregard any rational adult logic when their minds are made up.

Early the following morning, we packed our bags, checked out of the hotel, and drove to our final round of golf, which I had cleverly planned to be along our route back home. On the first tee, Brian pulled out his new, used prize possession and proceeded to hit a "duck hook," which is a golfer's term for a ball that goes very short and very left, often into a bunch of trees never to be seen again. I smiled nervously at Brian and tried to convince myself that he'd be fine on the next hole. After hitting duck hooks on holes two, three, and four, I found myself physically restraining Brian from throwing his new, used prized possession into the lake.

After our round was over, I drove back to Myrtle Beach to return the club, adding an hour to what would have been a 10-hour car ride. (I just hope Brian remembers times like these when I'm a frail old man drooling away in a retirement home.) At the golf store, the woman cheerfully said she would take the club back, but I needed to show her ... the receipt. Now I vaguely remembered putting the receipt someplace "special" just in case I would need it, but after packing, checking out, and playing golf, I would

134 Part 2: Probability Topics

have had a better chance of discovering a cure for cancer than remembering where I had put that piece of paper.

Not being one to give up easily, I marched back to the car and started unpacking everything. After a short while, during which time I had spread out my dirty underwear and socks all over their parking lot, the same woman walked out to tell me the store would gladly refund my money without the receipt if I would just pack up my things and put them back in the car.

I discovered a very powerful technique here that I am going to pass along to you. Just consider this a bonus for using my book. Whenever I can't find a receipt when I need to return something, I simply take along some dirty clothes in a suitcase and reenact my Myrtle Beach scenario right in front of the store. It works like a charm.

Anyway, let's assume that the number of tee shots that Brian normally hits that actually land in the fairway during a round of golf is five. The fairway is the area of short grass where the people who have designed this nerve-wracking game intended your tee shot to land. We will also assume that the actual number of fairways that Brian "hits" during one round follows the Poisson distribution.

.j^]— (Wrong Number

J^ Sp How do I know that the actual number of fairways that Brian "hits" during one

round follows the Poisson distribution? At this point, I really don't know for sure. What I would need to do to verify this claim is record the number of fairways hit over several rounds and then perform a "Goodness of Fit" test to decide whether the data fits the pattern of a Poisson distribution. I promise you that we will perform this test in Chapter 18, so please be patient.

We can now use the Poisson probability distribution to calculate the probability that Brian will hit x number of fairways during his next round, as follows:

x\ where:

x = the number of occurrences of interest over the interval H = the mean number of occurrences over the interval e = the mathematical constant 2.71828

P[x] = the probability of exactly x occurrences over the interval

We can now calculate the probability that Brian will hit exactly seven fairways during his next round. With H = S, the equation becomes this:

Bob's Basics



$P[7]:$

$$(5/7)(2.71838)^{-5} \cdot 1/7!$$

I



$$„m(78125)(0.006738)$$

$$P_7 = i^{-1} = 0.1044$$

$$7-6-5-4-3-2-1$$

In other words, Brian has slightly more than a 10 percent chance of hitting exactly seven fairways.

We can also calculate the cumulative probability that Brian will hit no more than two fairways using the following equations:

$$P[* < 2] = P[x = 0] + P[x = 1] + P[x = 2]$$

$$5^0)(2.71838)^{-5} \cdot 1/0.006738$$

$$P_{f*} = 0] = \wedge \wedge ^{-1} = \wedge ^{-1} = 0.0067$$

$$0! \cdot 1$$

$$(5^1)(2.71838)^{-5} \cdot (5)(0.006738)$$

$$P[x = 1] = L^A L = \backslash j y \cdot 1 = 0.0337$$

$$1! \cdot 1$$

$$(5^2)(2.71838)^{-5} \cdot (25)(0.006738)$$

$$P[X = 2] = \frac{e^{-\lambda} \lambda^2}{2!} = \frac{e^{-5} 5^2}{2!} = 0.0842$$

$$2! 2^{-1}$$

$$P[X \leq 2] = 0.0067 + 0.0337 + 0.0842 = 0.1246$$

Some statistics books use λ the symbol X , pronounced

lambda, to denote the mean of a Poisson probability distribution. However, regardless of the notation, it's still the same equation.

Bob's Basics

Remember from Chapter 8, If $0! = 1$. Also $x^0 = 1$ for any value of x .

There is a 12.46 percent chance that Brian will hit no more than two fairways during his next round.

In the previous example, the mean of the Poisson distribution happened to be an integer (5). However, this doesn't have to always be the case. Suppose the number of absent students for my Monday-morning statistics follows a Poisson distribution, with the average being 2.4 students. The probability that there will be three students absent next Monday is as follows.

136 Part 2: Probability Topics

$$P[X = 3] = \frac{e^{-\lambda} \lambda^3}{3!}$$

$$= \frac{e^{-2.4} (2.4)^3}{3!} = 0.090718$$

$$P[X = 3] = \frac{e^{-\lambda} \lambda^3}{3!} = 0.2090$$

$$3-2-1$$

Looks like I need to start taking roll on Mondays!

There's one more cool feature of the Poisson distribution: the variance of the distribution is the same as the mean. In other words:

This means that there are no nasty variance calculations like the ones we

dealt with in previous chapters for this distribution.

Poisson Probability Tables

Just like the binomial distribution, the Poisson probability distribution has a table that allows you to look up the probabilities for certain mean values. You can find the Poisson distribution table in Appendix B of this book. The following is an excerpt from this appendix with the probabilities from our Myrtle Beach example underlined.

Values of

u

x 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0

0 0.0408 0.0334 0.0273 0.0224 0.0183 0.0150 0.0123 0.0101 0.0082
0.0067

1 0.1304 0.1135 0.0984 0.0850 0.0733 0.0630 0.0540 0.0462 0.0395
0.0337

2 0.2087 0.1929 0.1771 0.1615 0.1465 0.1323 0.1188 0.1063 0.0948
0.0842

3 0.2226 0.2186 0.2125 0.2046 0.1954 0.1852 0.1743 0.1631 0.1517
0.1404

4 0.1781 0.1858 0.1912 0.1944 0.1954 0.1944 0.1917 0.1875 0.1820 0.1755

5 0.1140 0.1264 0.1377 0.1477 0.1563 0.1633 0.1687 0.1725 0.1747 0.1755

6 0.0608 0.0716 0.0826 0.0936 0.1042 0.1143 0.1237 0.1323 0.1398
0.1462

7 0.0278 0.0348 0.0425 0.0508 0.0595 0.0686 0.0778 0.0869 0.0959
0.1044

8 0.0111 0.0148 0.0191 0.0241 0.0298 0.0360 0.0428 0.0500 0.0575
0.0653

9 0.0040 0.0056 0.0076 0.0102 0.0132 0.0168 0.0209 0.0255 0.0307
0.0363

11 0.0004 0.0006 0.0009 0.0013 0.0019 0.0027 0.0037 0.0049 0.0064
0.0082

12 0.0001 0.0002 0.0003 0.0004 0.0006 0.0009 0.0013 0.0019 0.0026
0.0034

13 0.0000 0.0000 0.0001 0.0001 0.0002 0.0003 0.0005 0.0007 0.0009
0.0013

14 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0001 0.0002
0.0003 0.0005

15 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001
0.0001 0.0002

The probability table is organized by values of u , the average number of occurrences. Notice that the sum of each block of probabilities for a particular value of u . adds to 1.

As with the binomial tables, one limitation of using the Poisson tables is that you are restricted to using only the values of u . that are shown in the table. For instance, the previous table would not be useful for $u = 0.45$. However, other statistics books might contain Poisson tables that are more extensive than the one in Appendix B.

The Poisson distribution for $u = 5$ is shown graphically in the following histogram. The probabilities in Figure 10.1 are taken from the last column in the previous table.

.a ed .a

o

0.2

0.16

0.08

0.04

Figure 10.1

Poisson probability distribution.

4 6 8 10 12

Number of Occurrences

14

Note that the most likely number of occurrences for this distribution is four and five.

Here's another example. Let's assume that the number of car accidents each month at a busy intersection that I pass on my way to work follows the Poisson distribution with a mean of 1.8 accidents per month. What is the probability that three or more accidents will occur next month? You can express this as:

$$P[x > 3] = P[x = 3] + P[x = 4] + P[x = 5] + \dots + P[x = \infty]$$

138 Part 2: Probability Topics

Technically, with a Poisson distribution, there is no upper limit to the number of occurrences during the interval. You'll notice from the Poisson tables that the probability of a large number of occurrences is practically zero. Because we cannot add all the probabilities of an infinite number of occurrences (if you can, you're a much better statistician than I am!), we need to take 1 minus the complement of $P[x > 3]$ or:

$$P[x > 3] = 1 - P[x < 3] \text{ because:}$$

$$P[x = 0] + P[x = 1] + P[x = 2] + P[x = 3] + \dots + P[x = \infty] = 1 - 0$$

Therefore, to find the probability of three or more accidents, we'll use the following:

$$P[r > 3] = 1 - (P[* = 0] + P[r = i] + P[x = 2])$$

Using the probabilities underlined in the following Poisson table (I seem to have misplaced my calculator), we have this:

Values of λ

x 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0

0 0.3329 0.3012 0.2725 0.2466 0.2231 0.2019 0.1827 0.1653 0.1496
0.1353

1 0.3662 0.3614 0.3543 0.3452 0.3347 0.3230 0.3106 0.2975 0.2842
0.2707

2 0.2014 0.2169 0.2303 0.2417 0.2510 0.2584 0.2640 0.2678 0.2700
0.2707

3 0.0738 0.0867 0.0998 0.1128 0.1255 0.1378 0.1496 0.1607 0.1710
0.1804

4 0.0203 0.0260 0.0324 0.0395 0.0471 0.0551 0.0636 0.0723 0.0812
0.0902

5 0.0045 0.0062 0.0084 0.0111 0.0141 0.0176 0.0216 0.0260 0.0309
0.0361

6 0.0008 0.0012 0.0018 0.0026 0.0035 0.0047 0.0061 0.0078 0.0098
0.0120

7 0.0001 0.0002 0.0003 0.0005 0.0008 0.0011 0.0015 0.0020 0.0027
0.0034

8 0.0000 0.0000 0.0001 0.0001 0.0001 0.0002 0.0003 0.0005 0.0006
0.0009

9 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0001
0.0002

$$\sim P[x > 3] = 1 - (0.1653 + 0.2975 + 0.2678) \quad P[t > 3] = 1 - 0.7306 = 0.2694$$

There is almost a 27 percent chance that three or more accidents will occur in this intersection next month. Looks like I better find a safer way to work!

Using Excel to Calculate Poisson Probabilities

You can also conveniently calculate Poisson probabilities using Excel. The built-in POISSON function has the following characteristics:

POISSON(x, u, cumulative)

where:

cumulative = FALSE if you want the probability of exactly x occurrences

cumulative = TRUE if you want the probability of x or fewer occurrences

For instance, Figure 10.2 shows the POISSON function being used to calculate the probability that there will be exactly two accidents in the intersection next month.

Figure 10.2

POISSON function in Excel for exactly x occurrences.

Cell A1 contains the Excel formula =POISSON(2,1.8,FALSE) with the result being 0.2678. This probability is underlined in the previous table.

Excel will also calculate the cumulative probability that there will be no more than two accidents in the intersection, as shown in Figure 10.3.

Figure 10.3

POISSON function in Excel for no more than x occur-

Cell A1 contains the Excel formula =POISSON(2,1.8,TRUE) with the result being 0.7306, a probability that we saw in the last calculation and which is also the sum of the underlined probabilities in the previous table.

IM) Part 2: Probability Topics

One benefit of using Excel to determine Poisson probabilities is that you are not limited to the values of X shown in the Poisson table in Appendix B. Excel's POISSON function allows you to use any value for U .

Using the Poisson Distribution as an Approximation to the Binomial Distribution

I don't know about you, but when I have two ways to do something, I like to choose the one that's less work. If you don't agree with me, feel free to skip this material. If you do, read on!

We can use the Poisson distribution to calculate binomial probabilities under the following conditions:

- ◆ When the number of trials, n , is greater than or equal to 20 and ...
- ◆ When the probability of a success, p , is less than or equal to 0.05 ...

The Poisson formula would look like this:



Bob's Basics You might be asking yourself at this moment just why

you would want to do this. The answer is because the Poisson formula has fewer computations than the binomial formula and, under the stated conditions, the distributions are very close to one another.

If you need to calculate binomial probabilities with the number of trials, n , greater than or equal to 20 and the probability of a success, p , less than or equal to 0.05, you can use the equation for the Poisson distribution to approximate the binomial probabilities.

Just in case you are from Missouri (the "Show Me" state), I'll demonstrate this point with an example.

Suppose there are 20 traffic lights in my town and each has a 3 percent chance of not working properly

(a success) on any given day. What is the probability that exactly 1 of the 20 lights will not work today?

This is a binomial experiment with $n = 20$, $r = 1$, and $p = 0.03$. From Chapter 9, we

know that the binomial probability is this:

Chapter 10: The Poisson Probability Distribution HI

$$P[r, n] =$$

$$P[1, 20]:$$

$$(n-r)! \cdot p^r \cdot (1-p)^{n-r}$$

$$= 20! \cdot (0.03)^1 \cdot (0.97)^{19} \approx 0.3364$$

$$P[1, 20] = (20)(0.03)(0.97)^{19} = 0.3364$$

The Poisson approximation is as follows: $(np)^x \cdot e^{-np} / x!$ Because $np = (20)(0.03) = 0.6$:

$$(0.6)^1 \cdot e^{-0.6} / 1!$$

$$P[1] = \frac{(0.6)^1 \cdot e^{-0.6}}{1!}$$

$$1!$$

$$P[1] = (0.6)(0.548812) = 0.3293$$

Even if you're from Missouri, I think you would have to agree that the Poisson calculation is easier and the two results are very close. But if you need further proof .. Figures 10.4 and 10.5 show the histogram for each distribution for this example.

$$0.6 \quad 0.5$$

$$0.4 \quad 0.3 \quad 0.2 \quad 0.1$$

ii

Figure 10.4

The binomial probability distribution with $n = 20$, $p = 0.03$.

2 3 4

Number of Successes

1<i2 Part 2: Probability Topics

Figure 10.5

The Poisson probability distribution with the wean = 0.6.

2

(1.6 0.5 0.4 0.3 0.2 0,1

fl

0 12 3 4 5 6

Number or Occurrences

Even to a skeptic, these two distributions look very much alike. So my advice to you is to use the Poisson equation if you're faced with calculating binomial probabilities with $n > 20$ and $p < 0.05$.

This concludes our discussion of discrete probability distributions. I hope you've had as much fun with these as I've had!

Your Turn

1. The number of rainy days per month at a particular town follows a Poisson distribution with a mean value of six days. What is the probability that it will rain four days next month?
2. The number of customers arriving at a particular store follows a Poisson distribution with a mean value of 7.5 customers per hour. What

is the probability that five customers will arrive during the next hour?

3. The number of pieces of mail that I receive daily follows a Poisson distribution with a mean value of 4.2 per day. What is the probability that I will receive more than two pieces of mail tomorrow?

4. The number of employees who call in sick on Monday follows a Poisson distribution with a mean value of 3.6. What is the probability that no more than three employees will call in sick next Monday?

Chapter 10: The Poisson Probability Distribution

5. The number of spam e-mails that I receive each day follows a Poisson distribution with a mean value of 2.5. What is the probability that I will receive exactly one spam e-mail tomorrow?

6. Historical records show that 5 percent of people who visit a particular website purchase something. What is the probability that exactly 2 people out of the next 25 will purchase something? Use the Poisson distribution to estimate this binomial probability.

7. The number of times that Debbie proves me wrong each month follows a Poisson distribution with a mean of 2.5 times. What is the probability that "she who is never wrong" will fail to prove me wrong next month?

The Least You Need to Know

◆ A Poisson process counts the number of occurrences of an event over a period of time, area, distance, or any other type of measurement.

◆ The mean for a Poisson distribution is the average number of occurrences that would be expected over the unit of measurement and has to be the same for each interval of measurement.

◆ The number of occurrences during one interval of a Poisson process is independent of the number of occurrences in other intervals.

◆ If x is a Poisson random variable, the probability of x occurrences over the

interval of measurement is $pr^{\wedge}] _ E e \blacksquare$

$x \backslash$

◆ If the number of binomial trials is greater than or equal to 20 and the probability of a success is less than or equal to 0.05, you can use the equation for the Poisson distribution to approximate the binomial probabilities.

(hapteri

The Normal Probability Distribution

In This Chapter

- ◆ Examining the properties of a normal probability distribution
- ◆ Using the standard normal table to calculate probabilities of a normal random variable
- ◆ Using Excel to calculate normal probabilities
- ◆ Using the normal distribution as an approximation to the binomial distribution

Now let's take on a new challenge, continuous random variables and a continuous probability distribution known as the normal distribution. Remember that in Chapter 8 we defined a continuous random variable as one that can assume any numerical value within an interval as a result of measuring the outcome of an experiment. Some examples of continuous random variables are weight, distance, speed, or time.

The normal distribution is a statistician's workhorse. This distribution is the foundation for many types of inferential statistics that we rely on today. We will continue to refer to this distribution through many of the remaining chapters in this book.

IM> Part 2: Probability Topics

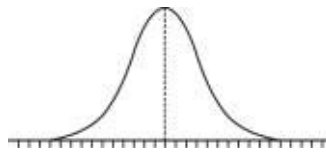
Characteristics of the Normal Probability Distribution

A continuous random variable that follows the normal probability distribution has several distinctive features. Let's say the monthly rainfall in inches for a particular city follows the normal distribution with an average of 3.5 inches and a standard deviation of 0.8 inches. The probability distribution for such a random variable is shown in Figure 11.1.

Figure 11.1

Normal distribution with a mean = 3.5, standard deviation = 0.8.

Normal Probability Distribution



0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 Mean = 3.5, Standard Deviation = 0.8

From this figure, we can make the following observations about the normal distribution:

- ◆ The mean, median, and mode are the same value—in this case, 3.5 inches.
- ◆ The distribution is bell-shaped and symmetrical around the mean.
- ◆ The total area under the curve is equal to 1.
- ◆ The left and right sides of the normal probability distribution extend indefinitely, never quite touching the horizontal axis.

The standard deviation plays an important role in the shape of the curve. Looking at the previous figure, we can see that nearly all the monthly rainfall measurements would fall between 1.0 and 6.0 inches. Now look at Figure 11.2, which shows the normal distribution with the same mean of 3.5 inches, but with a standard deviation of only 0.5 inches.

Normal Probability Distribution

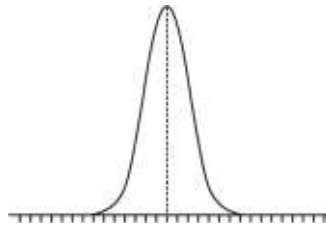


Figure 11.2

Normal distribution with a mean = 3.5, standard deviation = 0.5.

Mean = 3.5, Standard Deviation = 0.5

Here you see a curve that's much tighter around the mean. Almost all the rainfall measurements will be between 2.0 and 5.0 inches per month.

Figure 11.3 shows the impact of changing the mean of the distribution to 5.0 inches, leaving the standard deviation at 0.8 inches.

Bob's Basics



in

A smaller standard deviation results in a "skinnier" curve that's tighter and taller around the mean. A larger CT (standard deviation) makes for a "fatter" curve that's more spread out and not as tall.

Normal Probability Distribution

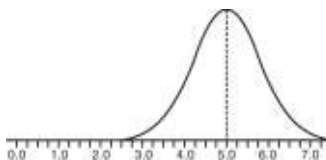


Figure 11.3

Normal distribution with a mean = 5.0, standard deviation = 0.8.

Mean = 5.0, Standard Deviation = 0.8

In each of the previous figures, the characteristics of the normal probability distribution hold true. In each case, the values of μ —the mean—and σ —the standard deviation—completely describe the shape of the distribution.

M Part 2: Probability Topics

The probability function for the normal distribution has a particularly mean personality (that pun was surely intended) and is shown as follows:

Good! I promise you this will be the last you'll see of this beast. Fortunately, we have other methods for calculating probabilities for this distribution that are more civilized and which we will discuss in the next section.

Calculating Probabilities for the Normal Distribution

There are a couple of approaches to calculate probabilities for a normal random variable. The following example demonstrates how this is done.

One morning a few days ago, Debbie called me on my cell phone while I was out running errands and spoke the two words that I had feared hearing for the past year. "They're back," she said. "Okay," I replied somberly, and then hung up the phone and headed straight toward the hardware store. My manhood was once again being challenged, and I'd be darned if I was going to take this lying down. This was war, and I was coining home fully prepared for battle! I am referring to, of course, my annual struggle with the most vile, the most dastardly, the most hungry creature that God has ever placed on this planet ... the Japanese beetle.

By the time I returned home from the hardware store, half of our beautiful plum tree looked like Swiss cheese. I quickly counterattacked with a vengeance, spraying the most potent chemicals money could buy. In the end, after the toxic spray cleared, I stood alone, master of my domain.

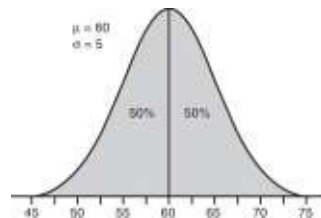
Alright, let's say that the amount of toxic spray I use each year follows a normal distribution with a mean of 60 ounces and a standard deviation of

5 ounces. This means that each year I do battle with these demons, the most likely amount of spray I'll use is 60 ounces, but it will vary year to year. The probability of other amounts above and below 60 ounces will drop off according to the bell-shaped curve. Armed with this information, we are now ready to determine probabilities of various usages each year.

Calculating the Standard Z-Score

Because the total area under a normal distribution curve equals 1 and the curve is symmetrical, we can say the probability that I will use 60 ounces or more of spray is 50 percent, as is the probability that I will use 60 ounces or less. This is shown in Figure 11.4.

Chapter 11: The Normal Probability Distribution 1M)



55 60 65 70

Ounces of Toxic Spray

Figure 1U

Normal distribution with a mean = 60, standard deviation = 5.

How would you calculate the probability that I would use 64.3 ounces of spray or less next year? I'm glad you asked. For this task, we need to define the standard normal distribution, which is a normal distribution with $\mu = 0$ and $\sigma = 1$, and is shown in Figure 11.5.

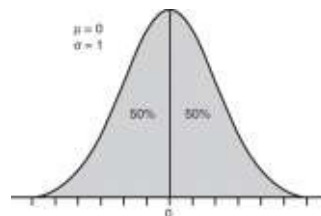


Figure 11.5

Standard normal distribution with a mean = 0, standard deviation = 1.0.

Number of Standard Deviations

This standard normal distribution is the basis for all normal probability calculations, and I'll use it throughout this chapter.

My next step is to determine how many standard deviations the value 64.3 is from the mean of 60 and show this value on the standard normal distribution curve. We do this using the following formula:

def

•i^ni'tien

The standard normal distribution is a normal distribution with a mean equal to 0 and a standard deviation equal to 1 .0.

-/'

ISO Part 2: Probability Topics

where:

X = the normally distributed random variable of interest

|J.= the mean of the normal distribution

CT= the standard deviation of the normal distribution

z = the number of standard deviations between x and μ , otherwise known as the standard z-score.

For the current example, the standard z-score is as follows:

$$\frac{64.3 - 60}{6.43} = 0.86$$

Now I know that 64.3 is 0.86 standard deviations away from 60 in my distribution.

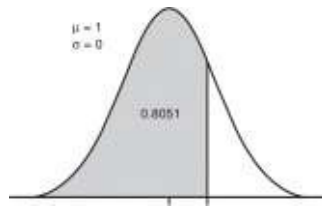
Using the Standard Normal Table

Now that I have my standard z-score, I can use the following table to determine the probability that I will use 64.3 ounces of toxic spray or less next year. This table is an excerpt from Appendix B and shows the area of the standard normal curve up to and including certain values of z. Because $z = 0.86$ in this example, we go to the 0.8 row and the 0.06 column to find a value of 0.8051, which is underlined.

Second digit of Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

This area is shown graphically in Figure 11.6.



0.86 Number of Standard Deviations

Figure 11.6

The shaded area represents the probability that z will be less than or equal to 0.86.

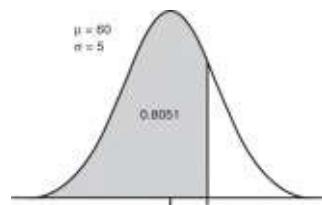
The probability that the standard z -score will be less than or equal to 0.86 is 80.51 percent. Because:

$$P[2 < 0.86] = P[x < 64.3] = 0.8051$$

There is an 80.51 percent chance I will use 64.3 ounces of spray or less next year against those evil Japanese beetles. This can be seen in Figure 11.7.

Figure 11.7

The shaded area represents the probability that x will be less than or equal to 64.3 ounces.



60 64.3 Ounces of Toxic Spray

152 Part 2: Probability Topics

\$

V^ Wrong Number

/* ■ - ^ P With continuous random variables, we cannot determine the probability of using exactly 64.3 ounces of spray because this would be an infinitely small probability. This is because I can use an infinite amount of quantities in any given year. One year, I could use 61.757 ounces and another year, 53.472 ounces. That's why with continuous random variables we can only calculate the probabilities of certain intervals, like less than 64.3 ounces or between 50.5 and 58.1 ounces. Compare this to discrete random variables from previous chapters. Because there were only a finite number of values for these variables, we could calculate the probability of exactly x occurrences or r successes.

What about the probability that I will use more than 62.5 ounces of spray next year? Because the standard normal table only has probabilities that are less than or equal to the z-scores, we need to look at the complement to this event.

$$P[Z > 62.5] = 1 - P[Z \leq 62.5]$$

The z-score now becomes this: $62.5 - 60$

$$= 0.50$$

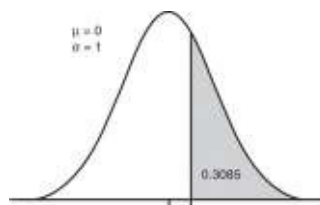
According to our normal table: $P[Z \leq 0.50] = 0.6915$

But we want:

$P[Z > 0.50] = 1 - 0.6915 = 0.3085$ This probability is shown graphically in Figure 11.8.

Figure 11.8

The shaded area represents the probability that z will be more than 0.50.



o 0.5 Number of Standard Deviations

Because:

$$P[z > 0.50] = P[x > 62.5] = 0.3085$$

There is a 30.85 percent chance that I will use more than 62.5 ounces of toxic spray. Beetles beware!

What about the probability that I will use more than 54 ounces of spray? Again, I need the complement, which would be this:

$$P[x > 54] = 1 - P[x < 54]$$

The z-score becomes this: $54 - 60$

-1.20

5

The negative score indicates that we are to the left of the distribution mean. Notice that the standard normal table only shows positive z values. But this is no problem because the distribution is symmetric. Figure 11.9 shows that the shaded area to the left of -1.2 standard deviations from the mean is the same as the shaded area to the right of +1.2 standard deviations from the mean.

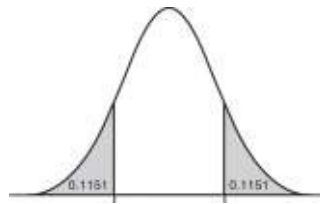


Figure 11.9

The shaded areas are equal.

-1.2 1.2

Number of Standard Deviations

We can determine the area to the right of + 1.2 standard deviations as follows: $P[z > +1.2] = 1 - P[z < +1.2] = 1 - 0.8849 = 0.1151$

Therefore, the area to the left of -1.2 standard deviations from the mean is also 0.1151. We now can calculate the area to the right of -1.2 standard deviations from the mean.

$$P[z > -1.2] = 1 - P[z < -1.2] = 1 - 0.1151 = 0.8849$$

\Vt Part 2: Probability Topics

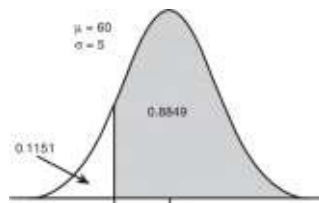
Because:

$$P[x > 54] = P[z > -1.2] = 0.8849$$

There is an 88.49 percent chance I will use more than 54 ounces of spray. This probability is shown graphically in Figure 11.10.

Figure 11.10

The shaded area is the probability that x will be more than 54 ounces.



54 60

Ounces of Toxic Spray

»>



Bob's Basics

A shortcut to the previous example would be to recognize the following:

$$P[z > -1.20] = P[z < +1.20]$$

$$P[z > -1.20] = 0.8849$$

In general, you can use the following two relationships for any value a

when dealing with negative z-scores:

$$P[z > -a] = P[z < +a]$$

$$P[z < -a] = 1 - P[z < +a]$$

Finally, let's look at the probability that I will use between 54 and 62.5 ounces of spray next year. This probability is shown graphically in Figure 11.11.

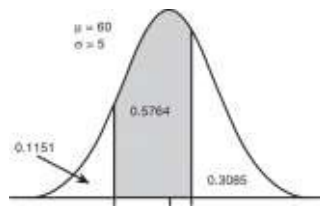


Figure 11.11

The shaded area is the probability that x will be between 54 and 62.5 ounces.

54 60 62.5

Ounces of Toxic Spray

We know from previous examples that the area to the left of 54 ounces is 0.1151 and that the area to the right of 62.5 ounces is 0.3085. Because the total area under the curve is 1:

$$P[54 < x < 62.5] = 1 - 0.1151 - 0.3085 = 0.5764$$

There is a 57.64 percent chance that I will use between 54 and 62.5 ounces of spray next year. I can't wait.

The Empirical Rule Revisited

Remember way back in Chapter 5 we discussed the empirical rule, which stated that if a distribution follows a bell-shaped, symmetrical curve centered around the mean, we would expect approximately 68, 95, and 99.7 percent of the values to fall within 1.0, 2.0, and 3.0 standard deviations around the mean respectively. I'm glad to inform you that we

now have the ability to demonstrate these results.

The shaded area in Figure 11.12 shows the percentage of observations that we would expect to fall within 1.0 standard deviation of the mean.

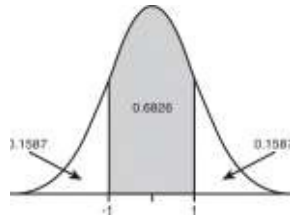


Figure 11.12

The shaded area is the probability that x will be between -1.0 and +1.0 standard deviation from the mean.

Number of Standard Deviations

1S6 Part 2: Probability Topics

Where did 68 percent come from? We can look in the normal table to get the probability that an observation will be less than one standard deviation from the mean.

$$P[z < +1.0] = 0.8413$$

Therefore, the area to the right of +1.0 standard deviations is this: $P[z > +1.0] = 1 - 0.8413 = 0.1587$

By symmetry, the area to the left of -1.0 standard deviations is also 0.1587. That leaves the area between -1.0 and +1.0 as this:

$$P[-1.0 < z < +1.0] = 1 - 0.1587 - 0.1587 = 0.6826$$

The same logic is used to demonstrate the probabilities of 2.0 and 3.0 standard deviations from the mean. I'll leave those for you to try.

Calculating Normal Probabilities Using Excel

Once again we can rely on Excel to do some of the grunt work for us. The first built-in function is NORMDIST, which has the following

characteristics:

`NORMDIST(^, mean, standard dev, cumulative)`

where:

cumulative = FALSE if you want the probability mass function (we don't)

cumulative = TRUE if you want the cumulative probability (we do)

Bob's Basics



Don't be alarmed if the values that are returned using the NORMDIST function in Excel are slightly different than those found in Table 3 in Appendix B. This is due to rounding differences that are small enough to be ignored.

For instance, Figure 11.13 shows the NORMDIST function being used to calculate the probability that I will use less than 64.3 ounces of spray on those nasty beetles next year.

Cell A1 contains the Excel formula `=NORMDIST (64.3,60,5,TRUE)` with the result being 0.8051. This probability is underlined in the previous table.

Excel also has a cool function called NORMSINV, which has the following characteristics:

`NORMSINV(probability)`

You provide this function a probability between 0 and 1, and it returns the corresponding z-score. Figure 11.14 shows the NORMSINV function returning a z-score for a probability of 0.8413, which is 1.0 standard deviation from the mean.

Figure 11.13

NORMDIST function in Excel for less than 64.3

Figure II.R

NORMSINV function in Excel for 1.0 standard deviation.

Cell A1 contains the Excel formula =NORMSINV(0.8413) with the result being 0.9998 (close enough to 1.0). If you look back to Figure 11.12, notice that the area to the left of 1.0 standard deviation from the mean totals to 0.8413. You can also find this value in the standard normal table next to $z = 1.0$.

Using the Normal Distribution as an Approximation to the Binomial Distribution

Remember how nasty our friend the binomial distribution can get sometimes? Well, the normal distribution may be able help us out during these difficult times under the right conditions. Recall from Chapter 9 that the binomial equation will calculate the probability of r successes in n trials with p - the probability of a success for each trial and q - the probability of a failure. If $np > 5$ and $nq > 5$, we can use the normal distribution to approximate the binomial.

158 Part 2: Probability Topics

As an example, suppose my statistics class is composed of 60 percent females. If I select 15 students at random, what is the probability that this group will include 8,

9, 10, or 11 female students? For this example, $n = 15$; $p = 0.6$; $q = 0.4$; and $r = 8, 9,$

10, and 11. We can use the normal approximation because $np = (15)(0.6) = 9$ and

$nq = (15)(0.4) = 6$. (Sorry, guys. I didn't mean to infer picking you would be classified a failure!)

^ Bob's Basics

k

Even though you are not interested in learning how the normal distribution can be used as an approximation to the binomial, I strongly encourage you to work through the example in this section. It will be good practice for determining probabilities for a normal distribution. And we all know that practice makes perfect!



Also recall from Chapter 9 that the mean and standard deviation of this binomial distribution is this:

$$\mu = np = (15)(0.6) = 9$$

$$\sigma = \sqrt{npq} = \sqrt{(15)(0.6)(0.4)} = 1.897$$

The probability that the group of 15 students will include 8, 9, 10, or 11 female students can be calculated using the following equations:

$$P[8,15] = {}^{15}C_8 (0.6)^8 (0.4)^7 = (6435)(0.0168)(0.0016) = 0.1730$$

$$P[9,15] = {}^{15}C_9 (0.6)^9 (0.4)^6 = (5005)(0.0101)(0.0041) = 0.2073$$

$$P[10,15] = {}^{15}C_{10} (0.6)^{10} (0.4)^5 = (3003)(0.0060)(0.0102) = 0.1838$$

$$P[11,15] = {}^{15}C_{11} (0.6)^{11} (0.4)^4 = (1365)(0.0036)(0.0256) = 0.1258$$

$$P[r = 8, 9, 10, \text{ or } 11] = 0.1730 + 0.2073 + 0.1838 + 0.1258 = 0.6899$$

Now let's solve this problem using the normal distribution and compare the results. Figure 11.15 shows the normal distribution with $\mu = 9$ and $\sigma = 1.897$.

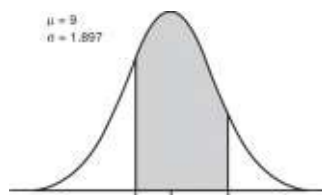


Figure 11.15

The normal approximation to the binomial distribution.

7.5 9 11.5

Number of Female Students

Notice that the shaded interval goes from 7.5 to 11.5 rather than 8 to 11. Don't worry; I didn't make a mistake. I subtracted 0.5 from 8 and added 0.5 to 11 to compensate for the fact that the normal distribution is continuous and the binomial is discrete. Adding and subtracting 0.5 is known as the continuity correction factor. For larger values of n , like 100 or more, you can ignore this correction factor.

Now we need to calculate the z-scores.

a

X-fi

$1 / \sqrt{npq}$

1.897

7.5-9

a

1.897

+1.32

-0.79

According to the normal table:

$P[z < +1.32] = 0.9066$ This area is shown in the shaded region of Figure 11.16.

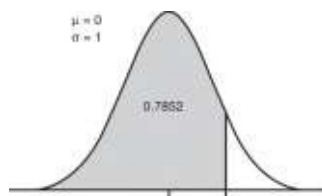


Figure 11.16

The probability that $z < +1.32$ standard deviations from, the mean.

0 +1.32

Number of Standard Deviations

160 Part 2: Probability Topics

We also know because of symmetry with the normal curve that: $P[z < -0.79] = 1 - P[z < +0.79]$

According to the normal table: $P[z < +0.79] = 0.7852$

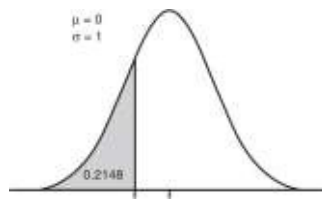
Therefore:

$$P[z < -0.79] = 1 - 0.7852 = 0.2148$$

This probability is shown in the shaded area in Figure 11.17.

Figure 11.17

The probability that $z < -0.79$ standard deviations from the mean.



-0.79 0 Number of Standard Deviations

The probability of interest for this example is the area between z-scores of -0.79 and +1.32. We can use the following calculations to find this area:

$$P[-0.79 < z < +1.32] = P[z < +1.32] - P[z < -0.79]$$

$$P[-0.79 < z < +1.32] = 0.9066 - 0.2148 = 0.6918$$

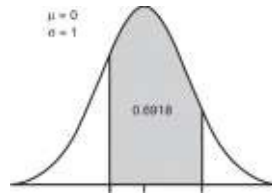
This probability is shown in the shaded area in Figure 11.18.

Figure 11.18

The probability that

$-0.79 < z < +1.32$ standard

deviations from, the mean.



-0.79 0 $+1.32$

Number of Standard Deviations

Using the normal distribution, we have determined the probability that my group of 15 students will contain 8, 9, 10, or 11 females is 0.6916. As you can see, this probability is very close to the result we obtained using the binomial equations, which was 0.6899.

Well, this ends our chapter on the normal probability distribution. And I feel much better prepared for next year's return visit of my archenemy, the Japanese beetle. Wish me luck.

Your Turn

1. The speed of cars passing through a checkpoint follows a normal distribution with $\mu = 62.6$ miles per hour and $\sigma = 3.7$ miles per hour. What is the probability that the next car passing will ...

- a. Be exceeding 65.5 miles per hour?
- b. Be exceeding 58.1 miles per hour?
- c. Be between 61 and 70 miles per hour?

2. The selling price of various homes in a community follows the normal distribution with $\mu = \$176,000$ and $\sigma = \$22,300$. What is the

probability that the next house will sell for ...

a. Less than \$190,000?

b. Less than \$158,000?

c. Between \$150,000 and \$168,000?

3. The age of customers for a particular retail store follows a normal distribution with $\mu = 37.5$ years and $\sigma = 7.6$ years. What is the probability that the next customer who enters the store will be ...

a. More than 31 years old?

b. Less than 42 years old?

c. Between 40 and 45 years old?

4. A coin is flipped 14 times. Use the normal approximation to the binomial distribution to calculate the probability of a total of 4, 5, or 6 heads. Compare this to the binomial probability.

162 Part 2: Probability Topics

5. A certain statistics author's golf scores follow the normal distribution with a mean of 92 and a standard deviation of 4. What is the probability that, during his next round of golf, his score will be ...

a. More than 97?

b. More than 90?

6. The number of text messages that Debbie's son Jeff sends and receives a month follows the normal distribution with a mean of 4,580 (I am not making this up!) and a standard deviation of 550. What is the probability that next month he will send and receive ...

a. Between 4,000 and 5,000 text messages?

b. Less than 4,200 text messages?

The Least You Need to Know

- ◆ The normal distribution is bell-shaped and symmetrical around the mean.
- ◆ The total area under the normal distribution curve is equal to 1.0.
- ◆ The normal distribution tables are based on the standard normal distribution where $\mu = 0$ and $\sigma = 1$.
- ◆ The number of standard deviations between a normally distributed random vari-

$X - \mu$

able, x , and σ is known as the standard z-score and can be found with $z = \frac{x - \mu}{\sigma}$.

a

- ◆ Excel has two built-in functions that you can use to perform normal distribution calculations: NORMDIST and NORMSINV.
- ◆ You can use the normal distribution to approximate the binomial distribution when $np > 5$ and $nq > 5$.

Part 3

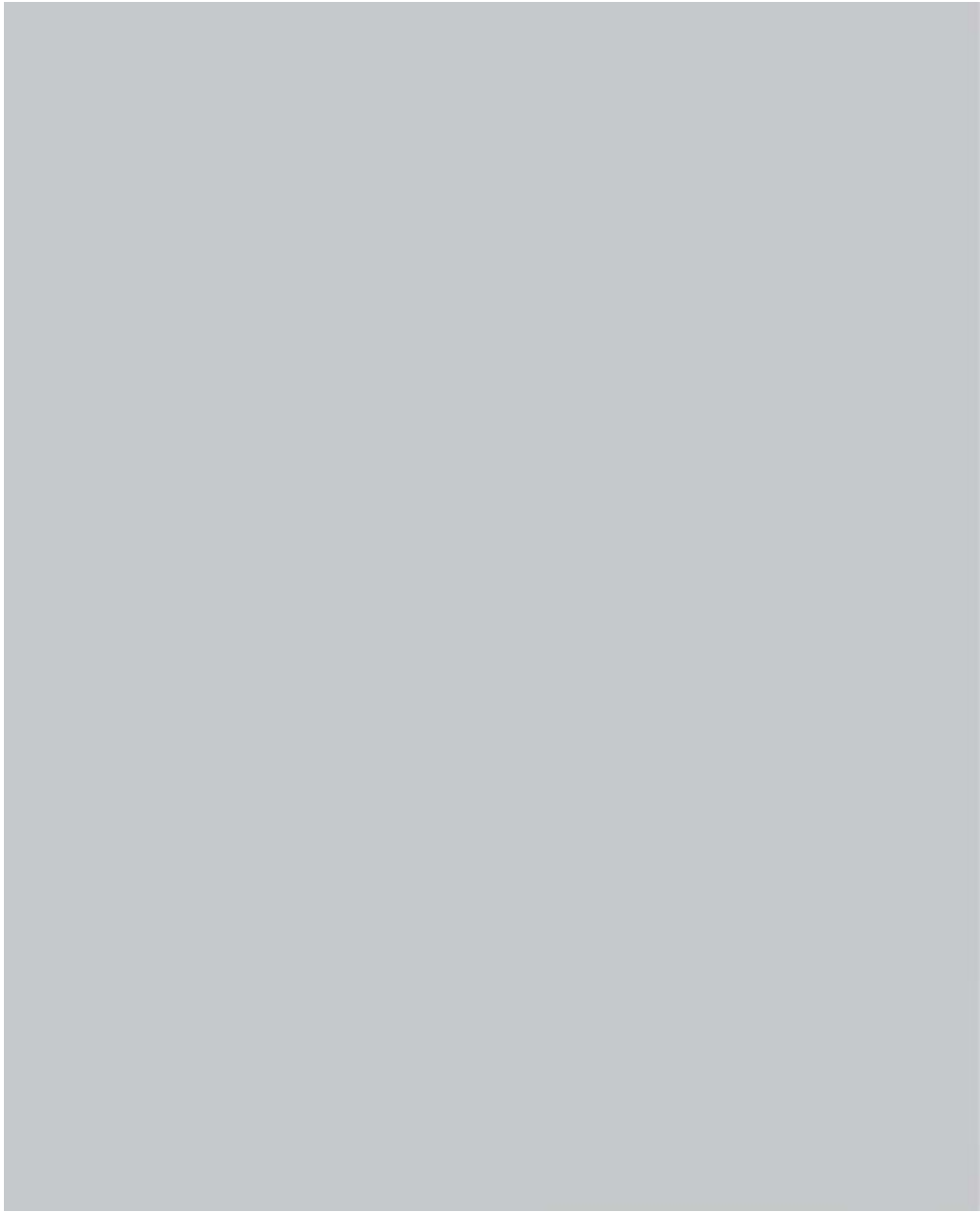
Inferential Statistics

Now we can take all those wonderful concepts that we have stuffed into our overloaded brains from Parts 1 and 2 and put them to work using statistically sounding words, such as confidence interval and hypothesis test. Inferential statistics enables us to make statements about a general population using the results of a random sample from that population. For instance, using inferential statistics, the winner of a political election

can be accurately predicted very early in the polling process based on the results of a relatively small random sample that is properly chosen. Pretty cool stuff!

X PIP A RANPOAA SAMPLING OF THE POPULATION ANP
PISCOVEREDTHAT78% LIKE MILK, 07% WATCH TELEV1SIOHAMP
100% HATE TO BE POLLEP.





(hapte

er
12

Sampling

In This Chapter

- ♦ The reason for measuring a sample rather than the population
- ♦ The various methods for collecting a random sample
- ♦ Defining sampling errors
- ♦ Consequences for poor sampling techniques

This first chapter dealing with the long-awaited topic of inferential statistics focuses on the subject of sampling. Way back in Chapter 1, we defined a population as representing all possible outcomes or measurements of interest, and a sample as a subset of a population. Here we'll talk about why we use samples in statistics and what can go wrong if they are not used properly.

Virtually all statistical results are based on the measurements of a sample drawn from a population. Major decisions are often made based on information from samples. For instance, the Nielson ratings gather information from a small sample of homes and are used to infer the television-viewing patterns of the entire country. The future of your favorite TV show rests in the hands of these select few! So choosing the proper sample is a critical step to ensure accurate statistical conclusions.

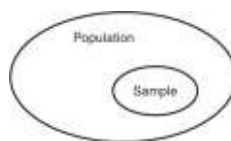
166 Part 3: Inferential Statistics

Why Sample?

Most statistical studies are based on a sample of the population at large. The relationship between a population and sample is shown in Figure 12.1 (and also described in Chapter 1).

Figure 12.1

The relationship between a population and sample.



Random Thoughts

Nielsen Media Research surveys 5,000 households nationwide to infer the television habits of millions of people. Because the results of these surveys are the basis for decisions such as show cancellations and advertising revenue, you better believe they select this sample very carefully.

Bob's Basics



Often it is just not feasible to measure an entire population. Even when it is feasible, measuring an entire population can be a waste of time and money and provides little added benefit beyond measuring a sample.

Why not just measure the whole population rather than rely on only a sample? That's a very good question! Depending on the study, measuring an entire population could be very expensive or just plain impossible. If I want to measure the life span of a certain breed of pesky mosquitoes (extremely short if I had any say in the matter), I could not possibly observe every single mosquito in the population. Rather, I would need to rely on a sample of the mosquito population, measure their life span, and make a statement about the life span of the entire population. That's the whole concept of inferential statistics in one paragraph! Unfortunately, doing what I just wrote is a whole lot harder than just writing it. Doing it is what the rest of this book is all about!

Even if we could feasibly measure the entire population, to do so would often be a wasteful decision. If a sample is collected properly and the analysis performed correctly, we can make a very accurate assessment of the entire population. There is very little added benefit to continue beyond the sample and measure everything in sight. Measuring the population often is a waste of both time and money, resources that seem to be very scarce these days.

Chapter 12: Sampling 167

One example where such a decision was recently made occurred at

Goldey-Beacom College, where I presently teach. I am also the Chair of the Academic Honor Code Committee and was involved in a project whose goal was to gather information regarding the attitude of our student body on the topic of academic integrity. It would have been possible to ask every student at our college to respond to the survey, but it was really unnecessary with the availability of inferential statistics. We eventually made the intelligent decision and sampled only a portion of the students to infer the attitudes of the population.

Random Sampling

The term random sampling refers to a sampling procedure where every member in the population has a chance of being selected. The objective of the sampling procedure is to ensure that the final sample to be measured is representative of the population from which it was taken. If this is not the case, then we have a biased sample, which can lead to misleading results. If you recall, we discussed an example of a biased sample back in Chapter 1 with the golf course survey. The selection of a proper sample is critical to the accuracy of the statistical analysis.

Since you can select a random sample in sev- If , •

aer # i # ni # tien

Random sampling refers to a sampling procedure where every member in the population has a chance of being selected. A biased sample sample that does not represent the intended population and can lead to distorted findings.

s a

eral ways, I'll use the following example to demonstrate these techniques.

Most of the time, I consider Debbie to be a person of sound mind and judgment (she married me, after all). Lately, however, I have had some concerns about her behavior dealing with the fact that she is reaching a major milestone in life before I am. Although I am not permitted to mention exactly what this milestone is (under penalty

of her not proofreading any more chapters and other certain activities), I

will say it involves dividing the number 100 by 2. (You do the math.)

Anyway, recently we were walking through the local mall when Debbie suddenly ran to a sales counter where they were selling fake ponytails for your hair. I had never heard of such a thing in my life and never would have conceived of the idea in a million years. Debbie, on the other hand, thought it was absolutely brilliant. Within seconds, a total stranger appeared from nowhere and before I could say, "That's my wife," had rearranged Debbie's hair and, in his final crowning moment, expertly arranged a fake hairpiece that somewhat resembled a small, furry animal on the top of her head.

168 PartJ: Inferential Statistics

Debbie, beaming with her "new look," turned to me to ask what I thought. Because this also happened to be our wedding anniversary, I weakly said it looked great as I handed this total stranger my credit card. (I might be a little slow in these matters, but I'm not stupid.) Debbie spent the rest of the evening prancing through the mall with her new cute furry animal hanging on for dear life. I have to admit, once I got used to the idea, it did look pretty cute.

Now let's say we wanted to conduct a survey to collect opinions of Debbie's new look. In fact, you, the reader, can render your opinion of Debbie after observing Figure 12.2 by sending me an e-mail from the book's website at www.stat-guide.com.

Figure 12.2

Debbie's new look; what do you think?



If I consider the current shoppers at the mall that night as my population, I need to decide how to select the random sample from whose opinion I will ask. As we will see in the following sections, there are four different ways to gather a random sample: simple random, systematic, cluster, and stratified.

Simple Random Sampling

A simple random sample is a sample in which every member of the population has an equal chance of being chosen. Unfortunately, this is easier said than done. In our mall example, I can randomly approach people to ask their opinion. However, I might have

Chapter 12: Sampling 169

some biases in my selection. For instance, if I observe that a certain menacing-looking person has a tattoo that says, "Death to All Statisticians," I might choose not to ask him what he thinks of Debbie's new ponytail. But in doing so, I might be biasing my sample.

Assuming I can rid myself of any biased selection, Figure 12.3 would describe a simple random sample at the mall.

defi'iortieio

A simple random sample is a

sample in which every member of the population has an equal chance of being chosen.

Figure I2.3

Simple random sample.

Each "X" represents a shopper, and each "X" that's circled represents a shopper in my sample.

There would be other options for choosing a simple random sample for the Academic Integrity survey mentioned earlier in the chapter. I could randomly choose students using a random number table, which is aptly named. (After all, it is simply a table of numbers that are completely random.) An excerpt of such a table is shown here:

57245 39666 18545 50534 57654 25519 35477 71309 12212 98911
42726 58321 59267 72742 53968 63679 54095 56563 09820 86291
82768 32694 62828 19097 09877 32093 23518 08654 64815 19894
97742 58918 33317 34192 06286 39824 74264 01941 95810 26247
48332 38634 20510 09198 56256 04431 22753 20944 95319 29515
26700 40484 28341 25428 08806 98858 04816 16317 94928 05512
66156 16407 57395 86230 47495 13908 97015 58225 82255 01956
64062 10061 01923 29260 32771 71002 58132 58646 69089 63694
24713 95591 26970 37647 26282 89759 69034 55281 64853 50837

170 Part 3: Inferential Statistics

M

Suppose we had 1,000 students in the population from which we were drawing a sample size of 100. (We'll discuss sample size in Chapter 14.) I would list these students with assigned numbers from 0 to 999. The random number table would tell me to select student 572, followed by student 427, and so forth until I had selected 100

students. Using this technique, my sample of 100 students would be chosen with complete randomness.

Random numbers can also be generated with Excel using the RAND function. Figure 12.4 demonstrates how this is done.

Cell A1 contains the formula =RAND(), which provides a random number between 0 and 1. This random number would result in student 357 being chosen for the sample.

de

Each time a change in the spreadsheet, Excel automatically recalculates all the functions and formulas, resulting in the generation of a new random number for each RAND function being used.

Figure 12.4

Excel's random number generator.

Systematic Sampling

definition

One way to avoid a personal bias when selecting people at random is to use systematic sampling. This technique results in selecting every k^{th} member of the population

to be in your sample. The value of k will depend on the size of the sample and the size of the population. Using my Academic Integrity survey, with a population of 1,000 students and a sample of 100, $k = 10$. From a listing of the entire population, I would choose every tenth student to be included in the sample. In general, if N = the size of the population and n - the size of the sample, the value of k

N would be approximately $\frac{N}{n}$.

In systematic sampling, every $(\frac{N}{n})^{\text{th}}$ member of the population is chosen for the sample, with the value of k being approximately $\frac{N}{n}$

Chapter 12: Sampling 171

We could also apply this sampling technique to the mall survey. Figure 12.5 shows every third customer walking into the mall being asked his or her opinion of Debbie's ponytail, even if the customer does have a tattoo.

Again, each "X" represents a shopper, and each "X" that's circled represents a shopper in my sample.

Figure 12.5

Systematic sampling.

©

The benefit of systematic sampling is that it's easier to conduct than a simple random sample, often resulting in less time and money. The downside is the danger of selecting a biased sample if there is a pattern in the population that is consistent with the value of k . For instance, let's say I'm conducting a survey on campus asking students how many hours they are studying during the week, and I select every fourth week to collect my data. Because we are on an 8-week semester schedule at Goldey-Beacom, every fourth week could end up being mid-terms and finals week, which would result in a higher number of study hours than normal (or at least I would hope so!).

Cluster Sampling

If we can divide the population into groups, or clusters, then we can select a simple random sample from these clusters to form the final sample. Using the Academic Integrity survey, the clusters could be defined as classes. We would randomly choose different classes to participate in the survey. In each class chosen, every student would be selected to be part of the sample.

172 Part 3: Inferential Statistics

def'i'ni'tietn

We could also conduct the mall survey using cluster sampling. Clusters

could be defined as stores in the mall population. We could randomly choose different stores and ask each customer in these stores his or her opinion about Debbie's ponytail. Figure 12.6 shows cluster sampling graphically.

Figure 12.6

Cluster sampling.

According to this figure, stores one, three, and four have been chosen to participate in the survey.

For cluster sampling to be effective, it is assumed that each cluster selected for the sample is representative of the population at large. In effect, each cluster is a miniaturized version of the overall population. If used properly, cluster sampling can be a very cost-effective way of collecting a random sample from the population. In the mall example, I would only have to visit three stores to conduct my survey, saving me valuable time on my wedding anniversary.

Stratified Sampling

In stratified sampling, we divide the population into mutually exclusive groups, or strata, and randomly sample from each of these groups. Using our mall example, we could define our strata as male and female shoppers. Using stratified sampling, I can be sure that my final sample contains an equal number of male and female shoppers. This can be shown graphically in Figure 12.7.

Chapter 12: Sampling 173

There are many different ways to establish strata from the population. Using the Academic Integrity survey, we could define our strata as undergraduate and graduate students. If 20 percent of our college population are graduate students, I could use stratified sampling to ensure that 20 percent of my final sample are also graduate students. Other examples of criteria that we can use to divide the population into strata are age, income, or occupation.

def<

i*r\i't\6r\

A stratified sample is obtained by dividing the population into mutually exclusive groups, or strata, and randomly sampling from each of these groups.

Figure I2.7

Stratified sampling.

Stratified sampling is helpful when it is important that the final sample has certain characteristics of the overall population. If we chose to use a simple random sample at the mall, the final sample may not have the desired proportion of males and females. This would lead to a biased sample if males feel differently about Debbie's new look than females.

Sampling Errors

Up to this point, we have stressed the benefits of drawing a sample from a population rather than measuring every member of the population. However, in statistics, as in life, there's no such thing as a free lunch. By relying on a sample, we expose ourselves to errors that can lead to inaccurate conclusions about the population.

Ylk Part 3: Inferential Statistics

The type of error that a statistician is most concerned about is called sampling error, which occurs when the sample measurement is different from the population measurement. Because the population is rarely measured in its entirety, the sampling error cannot be directly calculated. However, with inferential statistics, we'll be able to assign probabilities to certain amounts of sampling error later in Chapter 15.

Sampling errors occur because we might have the i\p(•! • y\'i • I 'iftr, unfortunate luck of selecting a sample that is not a

perfect match to the entire population. If the majority of mall shoppers really did like Debbie's new look but we just happened to choose a bunch of morons for our sample who did not fully appreciate a good thing when they saw it, Debbie might never wear her new ponytail again.

Sampling errors are expected and usually are a small price to pay to avoid measuring an entire population. One way to reduce the sampling error of a statistical study is to increase the size of the sample. In general, the larger the sample size, the smaller the sampling error. If you increase the sample size until it reaches the size of the population, then the sampling error will be reduced to zero. But in doing so, you forfeit the benefits of sampling.

Sampling error occurs when the sample measurement is different from the population measurement. It is the result of selecting a sample that is not a perfect match to the entire population.

Examples of Poor Sampling Techniques

The technique of sampling has been widely used, both properly and improperly, in the area of politics. One of the most famous mishaps with sampling occurred during the 1936 presidential race when the Literary Digest predicted Alf Landon to win the election over Franklin D. Roosevelt. Even if history is not your best subject, you can realize somebody had egg on his face after this election day. Literary Digest drew their sample from phonebooks and automobile registrations. The problem was that people with phones and cars in 1936 tended to be wealthier Republicans and were not representative of the entire voting population.

Another sampling blunder occurred in the 1948 presidential race when the Gallup poll predicted Thomas Dewey to be the winner over Harry Truman. The picture in Figure 12.8 shows a victorious Truman holding up the morning copy of the Chicago Tribune with the headline "Dewey Defeats Truman."

Chapter 12: Sampling 175



Figure 12.8

Dewey Defeats Truman.

The failure of the Gallup poll stemmed from the fact that there were a large number of undecided voters in the sample. It was wrongly assumed that these voters were representative of the decided voters who happened to favor Dewey. Truman easily won the election with 303 electoral votes compared to Dewey's 189.

*

\^, Wrong Number

^» // ^P Have you ever participated in an online survey on a sports or news website that allowed you to view the results? These surveys can be fun and interesting, but you need to take the results with a grain of salt. That's because the respondents are self-selected, which means the sample is not randomly chosen. The results of these surveys are most likely biased because the respondents would not be representative of the population at large. For example, people without Internet access would not be part of the sample and might respond differently than people with access to the Internet.

As you can see, choosing the proper sample is a critical step when using inferential statistics. Even a large sample size cannot hide the errors of choosing a sample that is not representative of the population at large. History has shown that large sample sizes are not needed to ensure accuracy. For example, the Gallup poll predicted that Richard Nixon would receive 43 percent of the votes for the 1968 presidential election and in fact he won 42.9 percent. This Gallup poll was based on a sample size of only 2,000; whereas the disastrous 1936 Literary Digest poll sampled 2,000,000 people (source: www.personal.psu.edu/faculty/g/e/gec7/Sampling.html).

176 Part 3: Inferential Statistics

Your Turn

You are to gather a systematic sample from a local phone book with

75,000 names. If every k^{th} name in the phone book is to be selected, what value of k would you choose to gather a sample size of 500?

Consider a population that is defined as every employee in a particular company. How could you use cluster sampling to gather a sample to participate in a survey involving employee satisfaction?

Consider a population that is defined as every employee in a particular company. How could you use stratified sampling to gather a sample to participate in a survey involving employee satisfaction?

The Least You Need to Know

- ◆ A simple random sample is a sample in which every member of the population has an equal chance of being chosen.
- ◆ In systematic sampling, every k^{th} member of the population is chosen for the

N sample, with the value of k being approximately $\frac{N}{n}$.

n

- ◆ A cluster sample is a simple random sample of groups, or clusters, of the population. Each member of the chosen clusters would be part of the final sample.
- ◆ Obtain a stratified sample by dividing the population into mutually exclusive groups, or strata, and randomly sampling from each of these groups.
- ◆ Sampling error occurs when the sample measurement is different from the population measurement. It is the result of selecting a sample that is not a perfect match to the entire population.

(hapter

13

Sampling Distributions

In This Chapter

- ◆ Using sampling distributions of the mean and proportion
- ◆ Working with the central limit theorem
- ◆ Using the standard error of the mean and proportion

In Chapter 12, we praised the wonders of using samples in our statistical analysis because it was more efficient than measuring an entire population. In this chapter, we'll discover another benefit of using samples—sampling distributions.

Sampling distributions describe how sample averages behave. You may be surprised to hear they behave very well—even better than the populations from which they are drawn. Good behavior means we can do a pretty good job at predicting future values of sample means with a little bit of information. This might sound a little puzzling now, but by the end of this chapter you'll be shaking your head in utter amazement.

What Is a Sampling Distribution?

Let's say I want to perform a study to determine the number of miles the average person drives a car in one day. Because it's not possible to measure the driving patterns of every person in the population, I randomly choose a

178 Part 3: Inferential Statistics

sample size of 10 ($n = 10$) qualified individuals and record how many miles they drove yesterday. I then choose another 10 drivers and record the same information. I do this three more times, with the results in the

following table.

Sample Number	Average Number of Miles (Sample Mean)
---------------	---------------------------------------

1	40.4
---	------

2	76.0
---	------

3	58.9
---	------

4	43.6
---	------

5	62.6
---	------

As you can see, each sample has its own mean value, and each value is different. We can continue this experiment by selecting many more samples and observe the pattern of sample means. This pattern of sample means represents the sampling distribution for the number of miles the average person drives in one day.

Sampling Distribution of the Mean

The distribution from the previous example represents the sampling distribution of the mean because the mean of each sample was the measurement of interest. This particular distribution has some interesting properties that I will discuss with the following example.

On a recent beach vacation, the resort we had chosen advertised a ping-pong tournament, which caught the eye of my 15-year-old son, John, who has enough skill to embarrass his poor old father. (This is the thanks I get for teaching him how to play the game when he had to stand on top of a cooler to see over the table.) As fate would have it, we were paired against each other and, to my relief, I found myself losing 10-8. With John needing one more point to win, I fed him two serves that he usually crams down my throat, but he somehow missed them both, and the score was tied at 10-10.

def # i*ni # tien

The sampling distribution of the mean refers to the pattern of sample

means that will occur as samples are drawn from the population at large.

If we had been playing back home in our basement, I'd have been dancing for joy and feeding him trash talk. But standing at that resort surrounded by spectators, all I could think about was a ruined vacation over a silly ping-pong game. On John's next serve, I attacked the ball with a motion that somewhat resembled a person having an epileptic seizure and hit the ball into the net. I gave my best "I can't believe I just

did that" expression and quickly sat down, breathing a hidden sigh of relief. But that was a small price to pay since John's pride was saved, as well as the rest of my vacation week. The things we do for our children!

Anyway, using ping-pong balls to describe the way sample means behave, assume that I have 100 ping-pong balls in a container in which 20 balls are marked with the number 1, 20 are marked with 2, 20 with 3, 20 with 4, and 20 with 5.

We can look at the probability distribution of this population in the following table.

Ball Number Frequency

Relative Frequency

Probability

20 20 20 20 20

20/100 20/100 20/100 20/100 20/100

0.20 0.20 0.20 0.20

0.20

This is known as a discrete uniform probability distribution because each event has the same probability, as you can see in Figure 13.1.

0.25

0.2

= 0.15 !5

CD .Q

2 0.1

Q.

0.05

Figure I3.1

Discrete uniform probability distribution.

2 3 4

Ball Number

180 Part 3: Inferential Statistics

def

•i # nrtien

A discrete uniform probability distribution is a distribution that assigns the same probability to each discrete event (and is discrete if it is countable).

We can calculate the mean and variance of a discrete uniform distribution as follows: 1,

$$-(a + b) r = -(b - a + 1)$$

where:

a = minimum value of the distribution b - maximum value of the distribution

For the ping-pong ball population: $H = 1(1+5) = 3.0$

a 2 =!(5-1 + 1) 2 :

12 v '

25 12

2.08

Keep these results in mind. We'll be referring to them later in the chapter.

Now for my demonstration. With the balls evenly mixed, I select one ball, record the number, place it back in the container, and then select a second ball, doing the same. This is my first sample with a size of 2 ($n = 2$). After doing this 25 times, I calculate the means of each sample and show the results in the following table.

Sampling Distribution of the Mean ($n = Z$)

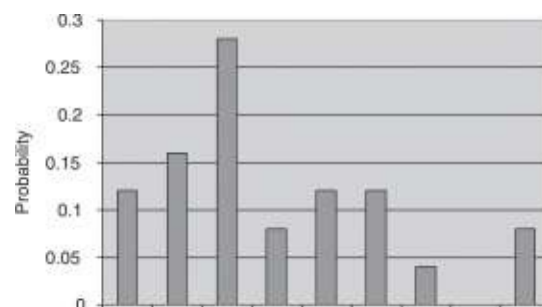
182 PartJ: Inferential Statistics

The previous table represents the sampling distribution of the mean for our ping-pong experiment with $n = 2$. We can show this distribution graphically in Figure 13.2.

I'm sure by now your highly inquisitive mind is screaming, "What happens to the sampling distribution if we increase the sample size?" That's an excellent question that I will address in the next section.

Figure 13.2

Sampling distribution of the mean for $n = 2$.



1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 Sample Means



Bob's Basics

The central limit theorem, in my humble opinion, is the most powerful concept for \inferential statistics. It forms the foundation for many statistical models that are used today It's a good idea to cozy up to this theorem.

The Central Limit Theorem

As I mentioned earlier, sample means behave in a very special way. According to the central limit theorem, as the sample size, n , gets larger, the sample means tend to follow a normal probability distribution. This holds true regardless of the distribution of the population from which the sample was drawn. Amazing, you say.

def<

i*r\i't\6r\

As you look at Figure 13.2, you're probably scratching your head and thinking, "That distribution doesn't look like a normal curve, which I know is bell-shaped and symmetrical." You're absolutely right because a sample size of two is generally not big enough for the central limit theorem to kick in.

Let's satisfy your curiosity and repeat my experiment by gathering 25 samples each consisting of 5 ping-pong balls ($n = 5$). I calculate the average of each sample and plot them in Figure 13.3.

Notice the impact that increasing the sample size has on the shape of the sample distribution. It's starting to appear somewhat bell-shaped with a little more symmetry. Let's look at sample sizes of 10 and 20 in Figures 13.4 and 13.5.

According to the central limit theorem, as the sample size, n , gets larger, the sample means tend to follow a normal probability distribution and tend to cluster around the true population mean. This holds true regardless of the distribution of the population from which the sample

was drawn.

0.45

0.4 0.35

0.3 0.25

0.2 0.15

0.1

0.05

0

Figure 13.3

Sampling distribution of the mean for $n = 5$.

1.0 1.5 2.0

2.5 3.0 3.5 Sample Means

4.0 4.5 5.0

18*1 PartJ: Inferential Statistics

Figure 13.4

Sampling distribution of the mean for $n = 10$.

0.6 0.5 0.4

-S 0.3

O

£ 0.2

0.1

1.0 1.5 2.0

2.5 3.0 3.5 Sample Means

4.0 4.5 5.0

1.0 1.5 2.0 2.5 3.0 3.5 Sample Means

4.0 4.5 5.0

Note that as the sample size increases, the sampling distribution tends to resemble a normal probability distribution. I don't know about you, but I find this pretty impressive considering the fact that the population that these samples were drawn from was not even close to being a normal distribution. If you recall, the ping-pong ball population followed a uniform distribution as shown in Figure 13.1.

Also, notice that as the sample size increases, the sample means tend to cluster around the true population mean, which if you recall we calculated as 3.0. This is another important feature of the central limit theorem.

And believe it or not, the central limit theorem has even one more important feature.

Standard Error of the Mean

Notice in the last four figures that as the sample size increased, the distribution of sample means tended to converge closer together. In other words, as the sample size increased, the standard deviation of the sample means became smaller. According to the central limit theorem (here we go again!), the standard deviation of the sample means can be calculated as follows:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

where:

$\sigma_{\bar{x}}$ = the standard deviation of the sample means σ = the standard deviation of the population n = sample size

The standard deviation of the sample means is formally known as the standard error of the mean.

Recall that earlier in the chapter, in the section "Sampling Distribution of the Mean," we determined that the variance of the ping-pong ball population was 2.08. Therefore:

$$s = \sqrt{\sigma^2/n} = \sqrt{2.08/2} = 1.02$$

We can now calculate the standard error of the mean for $n = 2$ in our example:

$$s_e = \frac{s}{\sqrt{n}} = \frac{1.02}{\sqrt{2}} = 0.72$$

4n - J2

defi'ni'ticm

The standard error of the mean

is the standard deviation of sample means. According to the central limit theorem, the standard error of the mean can

be determined by $s_e = \frac{s}{\sqrt{n}}$.



n

Bob's Basics.

Students often confuse σ and s . The symbol σ , the standard deviation of the population, measures the variation within the population and was discussed in Chapter 5. The symbol s , the standard error, measures the variation of the sample means and will decrease as the sample size increases.

186 Part 3: Inferential Statistics

The following table summarizes how the standard error varies with

sample size in our ping-pong ball example.

Standard Error Varies with Sample Size

Sample Size

Standard Error

2

5

10

20

1.02 0.64 0.46 0.32

Why Does the Central Limit Theorem Work?

Let me explain why the central limit theorem behaves the way it does. If this concept does not interest you, feel free to skip this section. I promise you won't hurt my feelings.

Going back to our original experiment with a sample size of two, the following table shows all the two-ball combinations that are possible along with the sample mean. This represents the theoretical sampling distribution of the mean because it represents all the possible combinations of samples along with their respective probabilities.

5 3.5

1 2.0

2 2.5

3 3.0

4 3.5

5 4.0

1 2.5

2 3.0

3 3.5

4 4.0

5 4.5

1 3.0

2 3.5

3 4.0

4 4.5

5 5.0

We can convert this table into a relative frequency distribution, which is shown in the following table.

Sample Mean x Frequency Relative Frequency Probability

1.0 1 $\frac{1}{25}$ 0.04

1.5 2 $\frac{2}{25}$ 0.08

2.0 3 $\frac{3}{25}$ 0.12

2.5 4 $\frac{4}{25}$ 0.16

3.0 5 $\frac{5}{25}$ 0.20

3.5 4 $\frac{4}{25}$ 0.16

4.0 3 $\frac{3}{25}$ 0.12

4.5 2 $\frac{2}{25}$ 0.08

5.0 1 1/25 0.04

This distribution is shown graphically in Figure 13.6.

188 PartJ: Inferential Statistics

def'i'ni'tiem

0.05

1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 Sample Means

You can see by this figure that the most common sample average is 3.0, whereas sample averages of 1.0 and 5.0 occur the least number of times. This is because there are more possible combinations of two-ball samples that average to 3.0 (5 to be exact) than two-ball samples that average to 1.0 or 5.0 (1 to be exact). In other words, we have five times the likelihood of drawing a two-ball sample that averages 3.0 when compared to sample averages of 1.0 or 5.0.

As we increase our sample size to 5, 10, and 20, the probability of drawing a sample with an average of 1.0 or 5.0 decreases while the probability of drawing a sample with an average of 3.0 increases. This explains why as sample size grows, more sample averages center around 3.0 and fewer around 1.0 and 5.0.

The theoretical sampling distribution of the mean displays all the possible sample means along with their classical probabilities. See Chapter 6 for a review of classical probability.

Putting the Central Limit Theorem to Work

I can just sense your need right now to do something really neat with this wonderful new tool. Look no further. If we know the sample means follow the normal probability distribution and we also know the mean and standard deviation of that distribution, we can predict the likelihood that the sample means will be greater or less than certain values.

For example, let's take our ping-pong ball experiment with $n = 20$. From the central limit theorem, we know the sample means follow a normal

distribution with:

$$\mu = 3.0$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 0.32$$

$$n = 20$$

What is the probability that our next sample of 20 ping-pong balls will have a sample average of 3.5 or less? The sample mean distribution is shown in Figure 13.7, with the shaded region indicating the probability of interest.

Sampling Distribution of the Mean

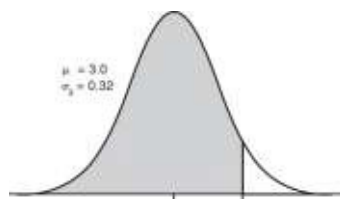
$$n = 20$$

Probability the next sample

mean will be less than or

equal to 3.5.

$$\mu = 3.0 \quad \sigma_{\bar{x}} = 0.32$$



$$3.0 \quad 3.5$$

Sample Means

As we did in Chapter 11, we need to calculate the z-score. The equation looks slightly different because we are working with sample means, but in reality, it is identical to what we saw in Chapter 11.

$$x - \mu,$$

$$z = \frac{3.5 - 3.0}{0.32} = 1.56$$

$$z = \frac{3.5 - 3.0}{0.32} = 1.56$$

$$z = 1.56$$

0.32 Using the standard z-table in Appendix B:

$P[\bar{x} < 3.5] = P[z < 1.56] = 0.9406$ This probability is shown in Figure 13.8.

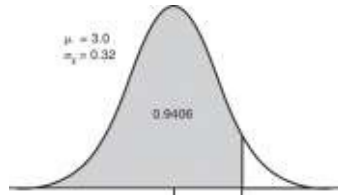
190 Part 3: Inferential Statistics

Figure 13.8

Probability the next sample mean will be less than or equal to 1.56 standard deviations from the population mean.

Sampling Distribution of the Mean $n = 20$

$$(i = 3.0 < \bar{x} \leq 3.5) = 0.32$$



$$0 + 1.56$$

Number of Standard Deviations

According to the shaded region, the probability that our next sample of 20 ping-pong balls will have a sample mean of 3.5 or less is approximately 94 percent.

The power of the central limit theorem lies in the fact that you need little information about the distribution of the population to apply it. The sample means will behave very nicely as long as the sample size is large enough. It's a very versatile theorem that has countless applications in the real world. I knew you'd be impressed!

Sampling Distribution of the Proportion

The sample mean is not the only statistical measurement that is performed. What if I want to measure the percentage of teenagers in this country who would agree with the following statement: "My parents are an excellent resource when I'm looking for advice on an important matter in my life." Because each respondent has only two choices (agree or disagree), this experiment follows the binomial probability distribution, which I discussed in Chapter 9.

Calculating the Sample Proportion

My measurement of interest is the proportion of teenagers in my sample of size n , who will agree with the statement "My parents are an excellent resource when I'm looking for advice on an important matter in my life." The sample proportion, p , is calculated by:

— Number of Successes in the Sample

$\frac{J}{n}$, Wrong Number

$\frac{1}{n}$

$\frac{1}{n}$

It's important to remember that a proportion, either p or p_s , cannot be less than 0 or greater than 1. A common mistake that students make is when told that the proportion equals 10 percent, they set $p = 10$ rather than $p = 0.1$.

Because I don't know the population proportion, p , who would agree with the statement, I need to collect data from samples and approximate the population proportion.

With proportion data, I want the sample size to be large enough so I can use the normal probability distribution to approximate the binomial distribution. As you recall from Chapter 11, if $np > 5$ and $nq > 5$, we can use the normal distribution to approximate the binomial ($q = 1 - p$, the probability of a failure). I'm hopeful that p will be at least 5 percent (at least a few teenagers might listen to their parents), so if I choose $n = 150$,

then:

$$B_p = (150)(0.05) = 7.5$$

$$b? = (150)(0.95) = 142.5$$

Suppose I choose 10 samples, each of size 150, and record the number of agreements (successes) in each sample in the table that follows.

Sample

Number of Successes

Sample Proportion

$$\begin{aligned} 26/150 &= 0.173 & 18/150 &= 0.120 & 21/150 &= 0.140 & 30/150 &= 0.200 & 24/150 &= \\ 0.160 & 21/150 &= 0.140 & 16/150 &= 0.107 & 28/150 &= 0.187 & 35/150 &= 0.233 \\ 27/150 &= 0.180 \end{aligned}$$

Next I average the sample proportions to approximate the population proportion, p : $0.173 + 0.12 + 0.14 + \dots + 0.233 + 0.18$

$P^*P, '-$

10

: 0.164

192 Part 3: Inferential Statistics

Calculating the Standard Error of the Proportion

I now need to calculate the standard deviation of this sampling distribution, which is known as the standard error of the proportion, or a, with the following equation:

$p\{i-p)$

$$0.164 \sqrt{1-0.164)}$$

150

defrnrteri

The standard error of the proportion is the standard deviation of the sample proportions and can be calculated by

$p(1-p)$.

$$\sqrt{0.000914} = 0.030$$

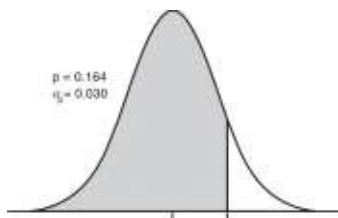
Now I'm ready to answer the age-old question, "What is the probability that from my next sample of 150 teenagers, 20 percent or less will agree with the statement: 'My parents are an excellent resource when I'm looking for advice on an important matter in my life'?" The shaded area in Figure 13.9 represents this probability, which displays the sampling distribution of the proportion for this example.

Figure 13.9

Sampling distribution of the proportion.

Sampling Distribution of the Proportion

$$p = 0.164 < r = 0.030$$



0.164 0.20 Sample Proportions

Because our sample size allows us to use the normal approximation to the binomial distribution, we now calculate the z-score for the proportion using the following equation:

P.-P

$$0.20 - 0.164$$

$$+1.20$$

0.030 Using the standard z-table in Appendix B:

$P[p \leq 0.20] = P[z < 1.20] = 0.8849$ This probability is also shown graphically in the shaded region in Figure 13.10.

Sampling Distribution of the Proportion

$p = 0.164$ $\alpha = 0.030$

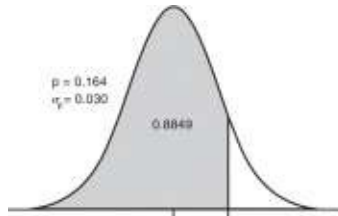


Figure 13.10

Probability that next sample proportion will be less than or equal to 1.2 standard deviations from, the population proportion.

0 +1.20

Number of Standard Deviations

According to our results, there is an 88.49 percent chance that 20 percent or fewer teenagers will agree with our statement from the next sample of size 150. Oh well, maybe when they get older, they'll discover the real wisdom of their parents.

Your Turn

1. Calculate the standard error of the mean when ...

a. $a = 10$, $n = 15$

b. $a = 4.7$, $n = 12$

c. $a = 7$, $n = 20$

2. A population has a mean value of 16.0 and a standard deviation of 7.5. Calculate the following with a sample size of 9.

a. $P[x < 17]$

194 Part 3: Inferential Statistics

b. $P[r > 18]$

c. $P[14.5 <^* < 16.5]$

3. Calculate the standard error of the proportion when ...

a. $p = 0.25, n = 200$

b. $p = 0.42, n = 100$

c. $p = 0.06, n = 175$

4. A population proportion has been estimated at 0.32. Calculate the following with a sample size of 160.

a. $P[p, < 0.30]$

b. $P[j_7, > 0.36]$

c. $P[0.29 \leq \text{ } \leq 0.37]$

5. A hypothetical statistics author is obsessed with making 10-foot putts. Each day that he practices, he putts 60 times and counts the number he makes. Over the last 20 practice sessions, he has averaged 24 made putts. What is the probability that he will make at least 30 putts during his next session?

The Least You Need to Know

◆ The sampling distribution of the mean refers to the pattern of sample means that will occur as samples are drawn from the population at large.

◆ According to the central limit theorem, as the sample size, n , gets larger, the sample means tend to follow a normal probability distribution.

◆ According to the central limit theorem, as the sample size, n , gets larger, the sample means tend to cluster around the true population

mean.

♦ The standard error of the mean is the standard deviation of sample means. According to the central limit theorem, the standard error of the mean can be

$\sigma_{\bar{x}}$

determined by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

* -in

♦ The standard error of the proportion is the standard deviation of the sample proportions and can be calculated by $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

(hapter

Confidence Intervals

14

In This Chapter

- ♦ Interpreting the meaning of a confidence interval
- ♦ Calculating the confidence interval for the mean with large and small samples
- ♦ Introducing the Student's t-distribution
- ♦ Calculating the confidence interval for the proportion
- ♦ Determining sample sizes to attain a specific margin of error

Now that we have learned how to collect a random sample and how sample means and sample proportions behave under certain conditions, we are ready to put those samples to work using confidence intervals.

One of the most important roles that statistics plays in today's world is to gather information from a sample and use that information to make a statement about the population from which it was chosen. We are using the sample as an estimate for the population. But just how good of an estimate is the sample providing us? The concept of confidence intervals will provide us with that answer.

196 Part 3: Inferential Statistics

Confidence Intervals for the Mean with Large Samples

So let's learn how to construct a confidence interval for a population mean using a large sample size. By a large sample size, we are generally referring to $n > 30$. The first step in developing a confidence interval for a population involves the following discussion on estimators.

Estimators

The simplest estimate of a population is the point estimate, the most common being the sample mean. A point estimate is a single value that best describes the population of interest. Let me explain this concept by using the following example.

I think my wife has been kidnapped and secretly Apt • | • y\\ • f Jam replaced by a Debbie look-alike who also happens to

be completely addicted to the QVC home shopping channel. No one or nothing in our household has escaped the products Debbie has found on her new favorite TV show. She has purchased stuff for the car, the kitchen floor, the dog, her skin, her hair, and my back (an inversion table that she wants me to hang upside down on!).

A point estimate is a single value that best describes the population of interest, the sample mean being the most common. An interval estimate provides a range of values that best describes the population.

Suddenly "Diamonique Week" has become a major holiday in our household. I'm not really sure what Diamonique actually is, but I suspect it is "available for a limited time only." Whenever I turn on any TV in the house, the channel always seems to be set to a very convincing home

shopping channel-type person pleading with me to "Call now! Only three left!"

Anyway, let's say I want to estimate the average dollar value of an order for the home shopping channel population. If my sample average was \$78.25, I could use that as my point estimate for the entire population of home shopping customers.

The advantage of a point estimate is that it is easy to calculate and easy to understand. The disadvantage, however, is that I have no clue as to how accurate this estimate really is.

To deal with this uncertainty, we can use an interval estimate, which provides a range of values that best describes the population. To develop an interval estimate, we need to learn about confidence levels.

Confidence Levels

A confidence level is the probability that the interval estimate will include the population parameter. A parameter is defined as a numerical description of a population characteristic, such as the mean.

Remember from Chapter 13 that sample means will follow the normal probability distribution for large sample sizes. Let's say we want to construct an interval estimate with a 90 percent confidence level. This confidence level corresponds to a z-score from the standard normal table equal to 1.64 as shown in Figure 14.1.

def'i'mi'ticm

A confidence level is the probability that the interval estimate will include the population parameter, such as the mean. A parameter is data that describes a characteristic about a population.

90% Confidence Interval

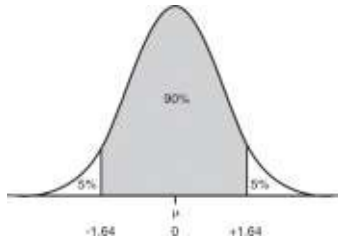


Figure R.1

90 percent confidence interval.

Bob's Basics.



Notice that in Figure 14.1, 5 percent of the area under the curve lies to the right of +1 .64 and 95 percent of the area under the curve lies to the left. That's why you see 0.9495 (close enough to 0.95) corresponding to a z-score of 1 .64 in Table 3 of Appendix B. Remember, however, that $z = 1 .64$ corresponds to a 90 percent confidence interval, the shaded region in the figure.

198 Part 3: Inferential Statistics

In general, we can construct a confidence interval around our sample mean using the following equations:

$\bar{x} + z_{\alpha/2} \cdot \sigma_{\bar{x}}$ (upper limit of confidence interval)

$\bar{x} - z_{\alpha/2} \cdot \sigma_{\bar{x}}$ (lower limit of confidence interval)

def

where:

\bar{X} = the sample mean

$z_{\alpha/2}$ - the critical z-score, which is the number of standard deviations based on the confidence level

$\sigma_{\bar{x}}$ = the standard error of the mean (remember our friend from Chapter

13?)

The term $z_{\alpha/2}$ is referred to as the margin of error, or E, a phrase often referred to in polls and surveys.

Going back to our home shopping example, let's say from a sample of 32 customers the average order is \$78.25 and the population standard deviation is \$37.50. (This represents the variation among orders within the population.) We can calculate our 90 percent confidence interval as follows:

$$\bar{x} = \$78.25$$

$$n = 32$$

$$s = \$37.50$$

• $z_{\alpha/2}$

A confidence interval is a range of values used to estimate a population parameter and is associated with a specific confidence level. The margin of error, E, determines the width of the confidence interval and is calculated using $z_{\alpha/2}$.

$z_{\alpha/2}$,

$$1.64$$

O-

• J_n

$$s = \$37.50$$

Si :

$$s = \$6.63$$

$$\text{Upper limit} = \bar{x} + z_{\alpha/2}s = \$78.25 + 1.64(\$6.63) = \$89.12$$

$$\text{Lower limit} = \bar{x} - z_{\alpha/2}s = \$78.25 - 1.64(\$6.63) = \$67.38$$

According to these results, our 90 percent confidence interval for this random sample of home shoppers is between \$67.38 and \$89.12 or (\$67.38, \$89.12). This interval is shown in Figure 14.2.

Interval Estimate for the Average Order Size of a Home Shopping Customer

\$67.38

\$78.25

Figure 14.2

Interval estimate for the average dollar value of a home shopping order.

Beware of the Interpretation of Confidence Interval!

As described earlier, a confidence interval is a range of values used to estimate a population parameter and is associated with a specific confidence level. A confidence interval needs to be described in the context of several samples. If we select 10 samples from our home shopping population and construct 90 percent confidence intervals around each of the sample means, then theoretically 9 of the 10 intervals will contain the true population mean, which remains unknown. Figure 14.3 shows this concept.

Interpretation of a 90% Confidence Interval

Figure 14.3

Interpreting the definition of a confidence interval.

This interval does not include μ

As you can see, Samples 1 through 9 have confidence intervals that include the true population mean, whereas Sample 10 does not.

μ is a Wrong Number

It is easy to misinterpret the definition of a confidence interval.

For example, it is not correct to state that "there is a 90 percent probability that the true population mean is within the interval (\$67.38, \$89.12)." Rather, a correct statement would be that "there is a 90 percent probability that any given confidence interval from a random sample will contain the true population mean."

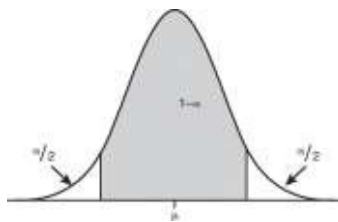
200 PartJ: Inferential Statistics

Because there is a 90 percent probability that any given confidence interval will contain the true population mean in the previous example, we have a 10 percent chance that it won't. This 10 percent value is known as the level of significance, α , which is represented by the total white area in both tails of Figure 14.4.

Figure R.A

The level of significance.

Level of Significance



def

•i mi»ticn

The level of significance (α) is the probability of making a Type I error.

The probability for the confidence interval is a complement to the significance level. For example, the significance level for a 95 percent confidence interval is 5 percent, the significance level for a 99 percent confidence interval is 1 percent, and so on. In general, a $(1 - \alpha)$ confidence interval has a significance level equal to α .

We will revisit the level of significance in more detail in later chapters.

The Effect of Changing Confidence Levels

So far, we have only referred to a 90 percent confidence interval. However, we can choose other confidence levels to suit our needs. The following table shows our home shopping example with confidence levels of 90, 95, and 99 percent.

Chapter R: Confidence Intervals ZOI

Confidence Intervals with Various Confidence Levels

Confidence Level

Sample Mean

Lower Limit

Upper Limit

From the previous table, you can see that there's a price to pay for increasing the confidence level—our interval estimate of the true population mean becomes wider and less precise. We have proven that once again, there is no free lunch with statistics. If we want more certainty that our confidence interval will contain the true population mean, that confidence interval will become wider.

Bob's Basics



/n

I recommend that you confirm the z-scores in this table for yourself by checking with Table 3 in Appendix B. Practice makes perfect! Review Chapter 11 if you need to.

The Effect of Changing Sample Size

There is one way, however, to reduce the width of our confidence interval while maintaining the same confidence level. We can do this by increasing the sample size. There is still no free lunch though because increasing the sample size has a cost associated with it. Let's say we increase our sample size to include 64 home shoppers. This change will affect our standard error as follows:

$q = \$37.50$ sin

a-

64

= \$4.69

Our new 90 percent confidence interval for our original sample will be: $\bar{x} = \$78.25$ $n = 64$ $a = \$4.69$ Upper limit = $\bar{x} + 1.64a = \$78.25 + 1.64(\$4.69) = \$85.94$

Lower limit = $\bar{x} - 1.64a = \$78.25 - 1.64(\$4.69) = \$70.56$

202 PartJ: Inferential Statistics

Increasing our sample size from 32 to 64 has reduced the 90 percent confidence interval from (\$67.38, \$89.12) to (\$70.56, \$85.94), which is a more precise interval.

Determining Sample Size for the Mean

We can also calculate a minimum sample size that would be needed to provide a specific margin of error. What sample size would we need for a 95 percent confidence interval that has a margin of error of \$8.00 ($E = \8.00) in our home shopping example?

$E = zC$

$E =$

$\sim E$

$(1.96)(\$37.50) \8.00

$= 84.4 \ll 85$

Therefore, to obtain a 95 percent confidence interval that ranges from

$\$78.25 - \$8.00 = \$70.25$ to $\$78.25 + \$8.00 = \$86.25$ would require a sample size

of 85 home shopping-addicted people.

Calculating a Confidence Interval When σ Is Unknown

Here's a simple section for you. (It's about time!) So far, all of our examples have assumed that we knew σ , the population standard deviation. What happens if σ is unknown? Don't panic, because as long as $n > 30$, we can substitute s , the sample standard deviation, for σ , the population standard deviation, and follow the same procedure as before. To demonstrate this technique, consider the following table that shows the order size in dollars of 30 home shoppers.

Home Shopping Sample [$n = 30$]

Using Excel, we can confirm that: $\bar{x} = \$84.47$ and $s = \$32.98$

A 99 percent confidence interval around this sample mean would be: $\bar{x} = \$84.47$, $n = 30$, $s = \$32.98$, $z_{\alpha/2} = 2.57$

$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\$32.98}{\sqrt{30}} = \$6.02$

We use $s_{\bar{x}}$ to indicate that we have approximated the standard error of the mean by using s instead of σ . We statisticians just love to put little hats on top of letters.

Upper limit = $\bar{x} + z_{\alpha/2} s_{\bar{x}} = \$84.47 + 2.57(\$6.02) = \99.94

Lower limit = $\bar{x} - z_{\alpha/2} s_{\bar{x}} = \$84.47 - 2.57(\$6.02) = \69.00

See! That wasn't too bad.

Using Excel's CONFIDENCE Function

Excel has a pretty cool built-in function that calculates confidence intervals for us. The CONFIDENCE function has the following characteristics:

CONFIDENCE(alpha, standard_dev, size)

where:

alpha = the significance level of the confidence interval
standard_dev =

the standard deviation of the population size = sample size

For instance, Figure 14.5 shows the CONFIDENCE function being used to calculate the confidence interval for our original home shopping example.

IQk Part 3: Inferential Statistics

Figure R.5

CONFIDENCE function in Excel for the home shopping sample.

Cell A1 contains the Excel formula =CONFIDENCE(0.1,37.5,32) with the result being 10.90394. This value represents the margin of error, or the amount to add and subtract from the sample mean, as follows:

$$\$78.25 + \$10.90 = \$89.15$$

$$\$78.25 - \$10.90 = \$67.35$$

This confidence interval is slightly different from the one calculated earlier in the chapter due to the rounding of numbers. This sure beats using tables and square root functions on the calculator.

Confidence Intervals for the Mean with Small Samples

So far, this entire chapter has dealt with the case where $n > 30$. I'm sure you are now wondering about how to construct a confidence interval when our sample size is less than 30. Well, as with many things in life, it depends.

With a small sample size, we lose the use of our faithful friend, the central limit theorem, and we need to assume that the population is normally (or approximately) distributed for all cases. The first case that we'll examine is when we know σ , the population standard deviation.

When σ Is Known

When σ is known, the procedure reverts back to the large sample size case. We can do this because we are now assuming the population is normally distributed. Let's construct a 95 percent confidence interval

from the following home shopping sample of size 10.

Home Shopping Sample [n = 10]

75

109 32

54

121 80

96

47

67

115

We know the following information: $\bar{x} = \$79.60$ $n = 10$ $\sigma = \$37.50$ (given from the original example)

$z_{\alpha/2} = 1.96$

$\sigma = \$37.50$, , , , , , $\sigma = ? = \sigma$; $\sigma^2 = \$11.86$ $\sigma = \sqrt{\$11.86}$

Upper Limit = $\bar{x} + z_{\alpha/2} \sigma = \$79.60 + 1.96 (\$11.86) = \102.85 Lower Limit = $\bar{x} - z_{\alpha/2} \sigma = \$79.60 - 1.96 (\$11.86) = \56.35

Notice that the small sample size has resulted in a wide confidence interval. Again, we are assuming here that the population from which the sample was drawn is normally distributed, which is the first time we have made such an assumption in this chapter so far.

When σ Is Unknown

More often, we don't know the value of σ . Here, we make a similar adjustment that we made earlier and substitute s , the sample standard deviation, for σ , the population standard deviation. However, because of the small sample size, this substitution forces us to use a new probability

distribution known as the Student's t-distribution (named in honor of you, the student).

Random Thoughts

The t-distribution is a continuous probability distribution with the following properties:

- ♦ It is bell-shaped and symmetrical around the mean.
- ♦ The shape of the curve depends on the degrees of freedom, (d.f.) which, when dealing with the sample mean, would be equal to $n - 1$.

The Student's t-distribution was developed by William Gosset (1876-1937) while working for the Guinness Brewing Company in Ireland. He published his findings using the pseudonym Student. Now there's a rare statistical event— a bashful Irishman!

206 Part 3: Inferential Statistics

def

•rrrticn

The degrees of freedom are the

number of values that are free to be varied given information, such as the sample mean, is known.

The area under the curve is equal to 1.0.

The t-distribution is flatter than the normal distribution. As the number of degrees of freedom increase, the shape of the t-distribution becomes similar to the normal distribution as seen in Figure 14.6. With more than 30 degrees of freedom (a sample size of 30 or more), the two distributions are practically identical.

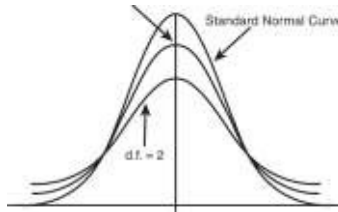
Figure R.6

The Student's t-distribution compared to the normal distribution.

The t-Distribution Compared to the Normal Curve

d.f. = 15

Standard Normal Curve



Students often struggle with the concept of degrees of freedom, which represent the number of remaining free choices you have after something has been decided, such as the sample mean. For example, if I know that my sample of size 3 has a mean of 10, I can only vary two values ($n - 1$). After I set those two values, I have no control over the third value because my sample average must be 10. For this sample, I have 2 degrees of freedom.

We can now set up our confidence intervals for the mean using a small sample:

$\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ (upper limit of confidence interval)

$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ (lower limit of confidence interval)

where:

critical t-value (can be found in Table 4 in Appendix B)

s

the estimated standard error of the mean

To demonstrate this procedure, let's assume the population of home shopping orders follows a normal distribution and the following sample of size 10 was collected.

Home Shopping Sample from a Normal Distribution ($n = 10$)

29

70

S9

100

4S

40

137

75

39

SS

With a unknown, we will construct a 95 percent confidence interval around the sample mean.

To determine the value of t_r for this example, I need to calculate the number

of degrees of freedom. Because $n = 10$, I have $n - 1 = 9$ d.f. This corresponds to

$t_r = 2.262$, which is underlined in the following table taken from Table 4 in Appendix

B.

Excerpt from the Student's t-Distribution

We next need to calculate the sample mean and sample standard deviation, which, according to Excel, are as follows:

$\bar{x} = \$71.50$ and $s^* = \$33.50$

208 PartJ: Inferential Statistics

We can now approximate the standard error of the mean:

$$s = \$33.50 \quad a = \frac{s}{\sqrt{n}} = \frac{\$33.50}{\sqrt{10}} = \$10.59$$

and can construct our 95 percent confidence interval:

$$\begin{aligned} \text{Upper limits } \bar{x} + t_{\alpha/2} \cdot s_{\bar{x}} &= \$71.50 + 2.262(\$10.59) = \$95.45 \\ \text{Lower limits } \bar{x} - t_{\alpha/2} \cdot s_{\bar{x}} &= \$71.50 - 2.262(\$10.59) = \$47.55 \end{aligned}$$

Now that wasn't too bad!

Bob's Basics



We can use the t-distribution when all of the following conditions have been met:

- ◆ The population follows the normal (or approximately normal] distribution.
- ◆ The sample size is less than 30.
- ◆ The population standard deviation, σ , is unknown and must be approximated by s , the sample standard deviation.

That ends our discussion on confidence intervals around the mean. Next on the menu are proportions!

Confidence Intervals for the Proportion with Large Samples

We can also estimate the proportion of a population by constructing a confidence interval from a sample. As you might recall from Chapter 13, proportion data follow the binomial distribution that can be approximated by the normal distribution under the following conditions:

$$np > 5 \text{ and } nq > 5$$

where:

p = the probability of a success in the population

Suppose I want to estimate the proportion of home shopping customers who are female based on the results of a sample. In Chapter 13, we learned that we can calculate the proportion of a sample using:

Number of Successes in the Sample

$P, = \sim$

n

Calculating the Confidence Interval for the Proportion

The confidence interval around the sample proportion can be calculated by: $p \pm Z \cdot \sigma_p$ (upper limit of confidence interval) $p - Z \cdot \sigma_p$ (lower limit of confidence interval)

where σ_p is the standard error of the proportion (which is the standard deviation of the sample proportions) using:

$P(1-P)$

There's extra credit for anyone who can see a problem arising here. Our challenge is that we are trying to estimate p , the population proportion, but we need a value for p to set up the confidence interval. Our solution—estimate the standard error by using the sample proportion as an approximation for the population proportion as follows:

We now can construct a confidence interval around the sample proportion by: $p \pm Z \cdot \sigma_p$ (upper limit of confidence interval) $p - z \cdot \sigma_p$ (lower limit of confidence interval)

Let's put these equations to work. In my efforts to estimate the proportion of female home shopping customers, I sample 175 random customers, of which 110 are female. I can now calculate p , the sample proportion:

Number of Successes in the Sample 110 __, __

$$p = \frac{S}{n} = \frac{110}{175} = 0.629$$

$$n = 175$$

The estimated standard error of the proportion would be:

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.629(1-0.629)}{175}} = 0.0365$$

210 Part 3: Inferential Statistics

We are now ready to construct a 90 percent confidence interval around our sample proportion ($z = 1.64$):

$$\begin{aligned} \text{Upper limit} &= p + z \cdot SE = 0.629 + 1.64(0.0365) = 0.689 \\ \text{Lower limit} &= p - z \cdot SE = 0.629 - 1.64(0.0365) = 0.569 \end{aligned}$$

Our 90 percent confidence interval for the proportion of female home shopping customers is (0.569, 0.689). Debbie must be in there somewhere!

Determining Sample Size for the Proportion

Almost done. Just as we did for the mean, we can determine a required sample size that would be needed to provide a specific margin of error. What sample size would we need for a 99 percent confidence interval that has a margin of error of 6 percent ($E = 0.06$) in our home shopping example? The formula to calculate n , the sample size is:

$$n = \frac{pq}{E^2}$$

Notice that we need a value for p and q . If we don't have a preliminary estimate of the values, set $p = q = 0.50$. Because half the population is female, that sounds like a good strategy to me.

$$z = 2.57$$

$$pq = (0.50)(0.50)$$

$$E = 0.06$$

:459

Therefore, to obtain a 99 percent confidence interval that provides a margin of error no more than 6 percent would require a sample size of 459 home shoppers.

Random Thoughts

The reason we use $p = q = 0.50$ if we don't have an estimate of the population proportion is that these values provide the largest sample size when compared to other combinations of p and q . It's like being penalized for not having specific information about your population. This way you are sure your sample size is large enough, regardless of the population proportion.

Your Turn

1. Construct a 97 percent confidence interval around a sample mean of 31.3 taken from a population that is not normally distributed with a standard deviation of 7.6 using a sample of size 40.
2. What sample size would be necessary to ensure a margin of error of 5 for a 98 percent confidence interval taken from a population that is not normal, which has a population standard deviation of 15?
3. Construct a 90 percent confidence interval around a sample mean of 16.3 taken from a population that is not normally distributed with a population standard deviation of 1.8 using a sample of size 10.
4. The following sample of size 30 was taken from a population that is not normally distributed:

10 4 9 12 5 17 20 9 4 IS

11 12 16 22 10 25 21 14 9 8 14 16 20 18 8 10 28 19 16 15

Construct a 90 percent confidence interval around the mean.

5. The following sample of size 12 was taken from a population that is normally distributed and that has a population standard deviation of

12.7:

37 48 30 55 50 46 40 62 50 43 36 66

Construct a 94 percent confidence interval around the mean.

6. The following sample of size 11 was taken from a population that is normally distributed:

121 136 102 115 126 106 115 132 125 108 130

Construct a 98 percent confidence interval around the mean.

7. The following sample of size 11 was taken from a population that is not normally distributed:

87 59 77 65 98 90 84 56 75 96 66 Construct a 99 percent confidence interval around the mean.

212 PartJ: Inferential Statistics

8. A sample of 200 light bulbs was tested, and it was found that 11 were defective. Calculate a 95 percent confidence interval around this sample proportion.

9. What sample size would you need to construct a 96 percent confidence interval around the proportion for voter turnout during the next election that would provide a margin of error of 4 percent? Assume the population proportion has been estimated at 55 percent.

The Least You Need to Know

A confidence interval is a range of values used to estimate a population parameter and is associated with a specific confidence level.

◆ A confidence level is the probability that the interval estimate will include the population parameter, such as the mean.

◆ Increasing the confidence level results in the confidence interval becoming wider and less precise.

- ◆ Increasing the sample size reduces the width of the confidence interval, which increases precision.
- ◆ Use the t-distribution to construct a confidence interval when the population follows the normal (or approximately normal) distribution, the sample size is less than 30, and the population standard deviation, σ , is unknown.
- ◆ Use the normal distribution to construct a confidence interval around the sample proportion when $np > 5$ and $nq > 5$.

(hapte

15

Introduction to Hypothesis Testing

In This Chapter

- ◆ Formulating the null and alternative hypothesis
- ◆ Distinguishing between a one-tail and two-tail hypothesis test
- ◆ Controlling the probability of a Type I and Type II error
- ◆ Determining the boundaries for the rejection region for the hypothesis test
- ◆ Stating the conclusion of the hypothesis test

Now that we know how to make an estimate of a population parameter, such as a mean, using a sample and a confidence interval, let's move on to the heart and soul of inferential statistics: hypothesis testing.

One thing statisticians like to do is to make a statement about a population parameter, collect a sample from that population, measure the sample, and declare, in a scholarly manner, whether or not the

sample supports the original statement. This, in a nutshell, is what hypothesis testing is all about. Of course, I've included a few juicy details. Without them, this would be one short chapter!

Part 3: Inferential Statistics

The purpose of this particular chapter is just to introduce the basic concept of hypothesis testing. The following two chapters will then get into more specific examples of how we put hypothesis testing to work. Stay tuned!

Hypothesis Testing—the Basics

In the statistical world, a hypothesis is an assumption about a population parameter. Examples of hypotheses (that's plural for hypothesis) include the following:

- ◆ The average adult drinks 1.7 cups of coffee per day.
- ◆ Twelve percent of undergraduate students will go directly to graduate school after graduation.
- ◆ No more than 2 percent of our products sold to customers are defective.

def •i'ni'tien

A hypothesis is an assumption about a population parameter.

In each case, we have made a statement about the population that may or may not be true. The purpose of hypothesis testing is to make a statistical conclusion about accepting or not accepting such statements. To further explain this concept, I present the following story.

I am man enough to admit that I am deathly afraid of snakes. That's why I did not hesitate to express my panic when Sam, Debbie's oldest teenage son, brought home a snake that he had caught and Debbie wholeheartedly agreed to let him keep in his bedroom.

Well, my worst nightmare came true the following morning. The snake had pushed off the top of the cage overnight and was loose somewhere in

the house. I guess Sam never heard the story of the mommy snake that once lifted a Volkswagen Beetle off her baby snake to save it.

I won't name names here, but smnehodys wife suggested that we put a mouse in Sam's room to attract the snake so we could catch it. I thought this was a very good joke until a white mouse showed up in Sam's room later that day posing as "snake bait."

That night, I lay in bed under high alert (i.e., at least one eye always open and ears finely tuned for a hissing noise) while Debbie lay calmly snoring next to me.

The next morning, I discovered that I had a new worst nightmare. The mouse had chewed its way out of its container overnight, and it, too, was loose somewhere in the house. I now had two wild animals roaming freely in the places where I eat, sleep, and

watch TV. By this time I'm frantically looking through the phonebook for a motel that specifically prohibits all snakes and mice. Debbie thought I was "overreacting."

That night I lay in my bed in the fetal position to protect my vital organs and keep my arms and legs away from the side of the bed while again Debbie lay calmly snoring next to me.

Anyway, let's try to tie this sci-fi tale to hypothesis testing. Let's say that my hypothesis is that it will take an average of six days to capture a loose snake in a house. In other words, I would like to test my belief that the population mean, μ , is equal to six days. I do this by gathering a sample of people who have had a loose snake in their home and calculate the average number of days required to capture it. Suppose the sample average is 6.1 days. The hypothesis test will then tell me whether or not 6.1 days is significantly different from 6.0 days or if the difference is merely due to chance. More details to follow!

The Null and Alternative Hypothesis

Every hypothesis test has both a null hypothesis and an alternative hypothesis. The null hypothesis, denoted by H_0 , represents the status quo and involves stating the belief that the mean of the population is $<$, $=$,

or $>$ a specific value. The null hypothesis is believed to be true unless there is overwhelming evidence to the contrary. In this example, my null hypothesis would be stated as:

$H_0 : \mu = 6.0 \text{ days}$

The alternative hypothesis, denoted by H_1 , represents the opposite of the null hypothesis and holds true if the null hypothesis is found to be false. The alternative hypothesis always states the mean of the population is $<$, $>$, or \neq a specific value. In this example, my alternative hypothesis would be stated as:

$H_1 : \mu \neq 6.0 \text{ days}$

The following table shows the three valid combinations of the null and alternative hypothesis.

df # i # r) i # tien

The null hypothesis, denoted by H_0 , represents the status quo and involves stating the belief that the mean of the population is $<$, $=$, or $>$ a specific value. The alternative hypothesis, denoted by H_1 , represents the opposite of the null hypothesis and holds true if the null hypothesis is found to be false.

216 Part 3: Inferential Statistics

Null Hypothesis

Alternative Hypothesis

$H_0 : \mu = 6.0$

$H_0 : \mu = 6.0$	$H_1 : \mu \neq 6.0$
$H_0 : \mu \geq 6.0$	$H_1 : \mu < 6.0$
$H_0 : \mu \leq 6.0$	$H_1 : \mu > 6.0$

Random Thoughts

Some textbooks will use the convention that the null hypothesis will always be stated as $=$ and will never use $<$ or $>$. Choosing either method

of stating your hypothesis will not affect the statistical analysis. Just be consistent with the convention you decide to use.

Note that the alternative hypothesis is never associated with $<$, proper combination is the topic of the next section.

or $>$. Selecting the

Stating the Null and Alternative Hypothesis

You need to be careful how you state the null and alternative hypothesis. Your choice will depend on the nature of the test and the motivation of the person conducting it.

If the purpose is to test that the population mean is equal to a specific value, such as our snake example, assign this statement as the null hypothesis, which results in the following:

$H_0 : \mu = 6.0 \text{ days}$

$H_1 : \mu \neq 6.0 \text{ days}$

Often hypothesis testing is performed by researchers who want to prove that their discovery is an improvement over current products or procedures. For example, if I invented a golf ball that I claimed would increase your distance off the tee by more than 20 yards, I would set up my hypothesis as follows:

$H_0 : \mu < 20 \text{ yards}$

$H_1 : \mu > 20 \text{ yards}$

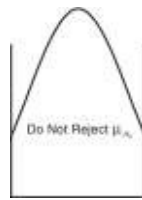
Note that I used the alternative hypothesis to represent the claim that I want to prove statistically so that I can make a fortune selling these balls to desperate golfers such as myself. Because of this, the alternative hypothesis is also known as the research hypothesis because it represents the position that the researcher wants to establish.

Two-Tail Hypothesis Test

A two-tail hypothesis test is used whenever the alternative hypothesis is expressed as \neq . Our snake example would involve a two-tail test because the alternative hypothesis is stated as $H_t: \mu \neq 6.0$. This test is shown graphically in Figure 15.1 which, as you can see, is considered a two-tail hypothesis test.

Two-Tail Hypothesis Test

Reject H_0



Reject H_0

6.0

Mean No. of Days to Catch a Snake

Figure 15.1

Two-tail hypothesis test.

def<

$\mu \neq 6.0$

The two-tail hypothesis test is

used whenever the alternative hypothesis is expressed as \neq .

The curve in the figure represents the sampling distribution of the mean for the number of days to catch a snake. The mean of the population, assumed to be 6.0 days according to the null hypothesis, is the mean of the sampling distribution and is designated by μ_0 .

The procedure is as follows:

◆ Collect a sample of size n , and calculate the test statistic, which in this case is the sample mean.

Plot the sample mean on the x-axis of the sampling distribution curve.

◆ If the sample mean falls within the white region, we do not reject H_0 . That is, we do not have enough evidence to support H_a , the alternative hypothesis, which states that the population mean is not equal to 6.0 days.

◆ If the sample mean falls in either shaded region, otherwise known as the rejection region, we reject H_0 . That is, we have enough evidence to support H_a which results in our belief that the true population mean is not equal to 6.0 days.

218 Part 3: Inferential Statistics

Because there are two rejection regions in this figure, we have a two-tail hypothesis test. We will discuss how to determine the boundaries for the rejection regions shortly.

H₀

, Wrong Number

$\alpha < \alpha/2$ The only two statements that we can make about the null hypothesis are that we ...

◆ Reject the null hypothesis.

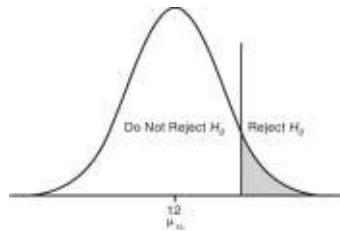
◆ Do not reject the null hypothesis. Because our conclusions are based on a sample, we will never have enough evidence to accept the null hypothesis. It's a much safer statement to say that we do not have enough evidence to reject H_0 . We can use the analogy of the legal system to explain. If a jury finds a defendant "not guilty," they are not saying the defendant is innocent. Rather, they are saying that there is not enough evidence to prove guilt.

One-Tail Hypothesis Test

A one-tail hypothesis test involves the alternative hypothesis being stated as $<$ or $>$. My golf ball example results in a one-tail test because the alternative hypothesis is being expressed as $H_a: \mu > 20$ and is shown in Figure 15.2.

Figure 15.2

One-tail hypothesis test.



Mean Increase in Yards off the Tee

def'i'ni'tiom

The one-tail hypothesis test

is used when the alternative hypothesis is being stated as \neq or $>$.

Here, there is only one rejection region, which is the shaded area on the right tail of the distribution. We follow the same procedure outlined for the two-tail test and plot the sample mean, which represents the average increase in distance from the tee with my new golf ball. Two possible scenarios exist.

If the sample mean falls within the white region, we do not reject H_0 . That is, we do not have enough evidence to support H_a the alternative hypothesis, which states that my golf ball increased distance off the tee by more than 20 yards. There goes my fortune down the drain!

If the sample mean falls in the rejection region, we reject H_0 . That is, we have enough evidence to support H_a which confirms my claim that my new golf ball will increase distance off the tee by more than 20 yards. Early retirement, here I come!

Bob's Basics



in

For a one-tail hypothesis test, the rejection region will always be consistent with the direction of the inequality for H_a . For $H_a: \mu > 20$, the rejection region will be in the right tail of the sampling distribution. For $H_a: \mu < 20$, the rejection region will be in the left tail.

Now that we have covered the basics of hypothesis testing, we need to consider errors that can occur due to sampling.

Type I and Type II Errors

Remember that the purpose of the hypothesis test is to verify the validity of a claim about a population based on a single sample. Because we are relying on a sample, we expose ourselves to the risk that our conclusions about the population will be wrong.

Using the golf ball example, suppose that my sample falls within the "Reject H_0 " region of the last figure. That is, according to the sample, my golf ball increases distance off the tee by more than 20 yards. But what if the true population mean is actually much less than 20 yards? This can occur primarily because of sampling error, which I discussed in Chapter 12. This type of error, when we reject H_0 when in reality it's true, is known as a Type I error. The probability of making a Type I error is known as α , the level of significance, which I first introduced in Chapter 14.

We also can experience another type of error with hypothesis testing. Let's say the golf ball sample fell within the "Do Not Reject H_0 " region of the last figure. That is, according to the sample, my golf ball does not increase the distance off the tee by more than 20 yards. But what if the true population mean is actually much more than 20 yards? This type of error, when we do not reject H_0 when in reality it's false, is known as a Type II error. The probability of making a Type II error is known as β .

The following table summarizes the two types of hypothesis errors.

H_0 Is True

H_0 Is False

Reject H_0

Do Not Reject H_0

Type I Error $P[\text{Type I Error}] = \alpha$ Correct Outcome

Correct Outcome

Type II Error $P[\text{Type II Error}] = \beta$

definition

A Type I error occurs when the null hypothesis is not accepted when in reality it is true. A Type II error occurs when we fail to reject the null hypothesis when in reality it is not true.

Normally, with hypothesis testing, we decide on a value for α that is somewhere between 0.01 and 0.10 before we collect the sample. The value of β can then be calculated, but that topic goes beyond the scope of this book. Be grateful for this because that concept is very complicated!

Let's put these concepts to work now and do some real hypothesis testing!

Random Thoughts

Ideally, we would like the values of α and β to be as small as possible. However, for a given sample size, reducing the value of α will result in an increase in the value of β . The opposite also holds true. The only way to reduce both α and β simultaneously is to increase the sample size. Once the sample size has been increased to the size of the population, the values of α and β will be 0. However, as we discussed in Chapter 12, this is not a recommended strategy.

Example of a Two-Tail Hypothesis Test

I stated the hypotheses for the snake example as: $H_0: \mu = 6.0 \text{ days}$ $H_1: \mu \neq 6.0 \text{ days}$

Where μ = the mean number of days to catch a loose snake in a home.

Let's say that I know that the standard deviation of the population, σ , is 0.5 days, and my sample size to test the hypothesis, n , is 30 homes. (Please don't ask me how I'm going to find 30 homes with loose snakes. I'm making this up as I go along, so just humor me.) We'll also set $\alpha = 0.05$, which means I'm willing to accept a 5 percent chance of committing a Type I error. Our first step is to calculate the standard error of the mean, $\sigma_{\bar{x}}$. If you remember from Chapter 13, the equation is:

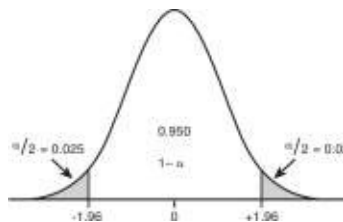
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{30}} = 0.0913$$

Let's assume the sample mean from the 30 homes is 6.1 days. What is our conclusion about our estimate of the population mean, μ ?

To answer this, we next have to determine the critical z-score, which corresponds to $\alpha = 0.05$. Because this is a two-tail test, this area needs to be evenly divided between both tails, with each tail receiving $\alpha/2 = 0.025$. According to Figure 15.3, we need to find the critical z-score that corresponds to the area $0.950 + 0.025 = 0.975$. As you can see, the 0.950 area is derived from $1 - \alpha$.

Figure 15.3

Critical z-score for $\alpha = 0.05$.



$$\alpha/2 = 0.025$$

Using Table 3 in Appendix B, we look for the closest value to 0.9750 in the body of the table. We can find this value by looking across column 1.9 and down row 0.06 to arrive at the z-score of +1.96 for the right tail and

-1.96 for the left tail.

Using the Scale of the Original Variable

Now let's determine the rejection region using the scale of the original variable, which in this case is the number of days. To calculate the upper and lower limits of the rejection region, we use the following equations. Recall from Chapter 14 that we use the z-scores from the standard normal distribution when $n > 30$ and CT is known.

Limits of rejection region = $\mu_H + z_{\alpha/2} \sigma_{\bar{X}}$

222 PartJ: Inferential Statistics

where μ_H = the population mean assumed by the null hypothesis.

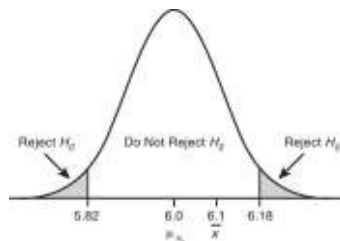
For our snake example:

Upper limit = $\mu_H + z_{\alpha/2} \sigma_{\bar{X}} = 6.0 + (1.96)(0.0913) = 6.18$ days
Lower limit = $\mu_H - z_{\alpha/2} \sigma_{\bar{X}} = 6.0 + (-1.96)(0.0913) = 5.82$ days

Because our sample mean is 6.1 days, this falls within the "Do Not Reject H_0 " region as shown in Figure 15.4. Our conclusion is that the difference between 6.1 days and 6.0 days is merely due to chance variation, and we have support that the population mean is 6 days.

Figure 15.4

Hypothesis test for the snake example (original variable scale).



Mean Number of Days to Catch a Snake

Using the Standardized Normal Scale

We can arrive at the same conclusion by setting up the boundaries for the rejection region using the standardized normal scale. We do this by calculating the z-score that corresponds to the sample mean as follows:

-»h.

6.1-6.0 0.0913

hi.09

h



Bob's Basics

Be sure to distinguish between the calculated z-score and the critical z-score. The calculated z-score, z , represents the number of standard deviations between the sample mean and μ_H , the population mean according to the null hypothesis. The critical z-score, z_{α} , is based on the significance level, α , and determines the boundary for the rejection region.

Figure 15.5 shows this result graphically. Because the calculated z-score of +1.09 is within the "Do Not Reject H_0 " region, the conclusions of both techniques are consistent.

Number of Standard Deviations from the Mean

Example of a One-Tail Hypothesis Test

Because I formulated the alternative hypothesis for the golf ball example as > 20 , this becomes a one-tail test. The hypothesis for this example is stated as:

$H_0 : \mu \leq 20$ yards

$H_1 : \mu > 20$ yards

Where μ - the mean increase in yards off the tee using my new golf ball.

Let's say that I know that the standard deviation of the population, σ , is 5.3 yards and my sample size to test the hypothesis, n , is 40 golfers. For this example, we'll set $\alpha = 0.01$. The standard error of the mean, $\sigma_{\bar{x}}$, will now be equal to:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5.3}{\sqrt{40}}$$

$$\sigma_{\bar{x}} = 0.838$$

Let's assume the sample mean from the 40 golfers is 22.5 yards. What is our conclusion about our estimate of the population mean, μ ?

Once again, we next have to determine the critical z-score, which corresponds to $\alpha = 0.01$. Because this is a one-tail test, this entire area needs to be in one rejection region on the right side of the distribution. According to Figure 15.6, we need to find the z-score that corresponds to the area 0.99 or $1 - \alpha$.

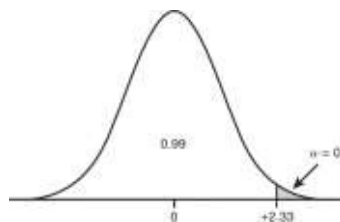
$$z_{\alpha} = 2.33$$

Part 3: Inferential Statistics

Using Table 3 in Appendix B, we look for the closest value to 0.9900 in the body of the table, which results in a critical z-score of 2.33.

Figure 15.6

Critical z-score for $\alpha = 0.01$.



Number of Standard Deviations from the Mean

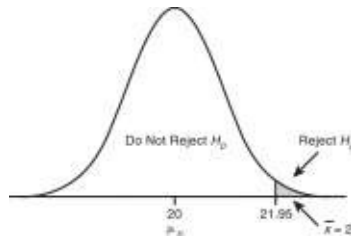
To calculate the limit for this rejection region using the scale of the original variable, we use:

$$\text{Limit} = \bar{x} + z_{\alpha} \sigma_{\bar{x}} = 20 + (2.33)(0.838) = 21.95 \text{ yards}$$

Because our sample mean is 22.5 yards, this falls within the "Reject H_0 " region as shown in Figure 15.7. Our conclusion is that we have enough evidence to support the hypothesis that the mean increase in distance off the tee with my new balls exceeds 20 yards. I'm in business!

Figure 15.7

Hypothesis test for the golf ball example (original variable scale).



Mean Increase in Distance off the Tee in Yards

Random Thoughts

You might be asking yourself, "If the sample mean was 21.0 yards, shouldn't that provide conclusive evidence that the new ball increases distance by more than 20 yards?" According to the previous figure, the answer is no. Because we are basing our decision on a sample, an average of 21 is just too close to 20 to satisfy my claim. The sample average would have to be 21.95 yards or more in order to reject the null hypothesis.

As I mentioned earlier, the purpose of this chapter was to introduce the basic concepts of hypothesis testing. The following two chapters will explore hypothesis testing in even more loving detail. So hang in there—we're just getting warmed up!

Your Turn

1. Formulate a hypothesis statement for the following claim: "The average adult drinks 1.7 cups of coffee per day." A sample of 35 adults drank an average of 1.95 cups per day. Assume the population standard deviation is 0.5 cups. Using $\alpha = 0.10$, test your hypothesis. What is your conclusion?
2. Formulate a hypothesis statement for the following claim: "The average age of our customers is less than 40 years old." A sample of 50

customers had an average age of 38.7 years. Assume the population standard deviation is 12.5 years. Using $\alpha = 0.05$, test your hypothesis. What is your conclusion?

3. Formulate a hypothesis statement for the following claim: "The average life of our light bulbs is more than 1,000 hours." A sample of 32 light bulbs had an average life of 1,190 hours. Assume the population standard deviation is 325 hours. Using $\alpha = 0.02$, test your hypothesis. What is your conclusion?

4. Formulate a hypothesis statement for the following claim: "The average delivery time is less than 30 minutes." A sample of 42 deliveries had an average time of 26.9 minutes. Assume the population standard deviation is 8 minutes. Using $\alpha = 0.01$, test your hypothesis. What is your conclusion?

5. Formulate a hypothesis statement for the following claim: "Students graduating from college have an average credit card debt of \$2,700." A sample of 40 college graduates averaged \$2,450 in credit card debt. Assume the population standard deviation is \$950. Using $\alpha = 0.05$, test your hypothesis. What is your conclusion?

226 Part 3: Inferential Statistics

The Least You Need to Know

◆ The null hypothesis, denoted by H_0 , represents the status quo and involves stating the belief that the mean of the population is $<$, $=$, or $>$ a specific value.

◆ The alternative hypothesis, denoted by H_a represents the opposite of the null hypothesis and holds true if the null hypothesis is found to be false.

◆ Use a two-tail hypothesis test whenever the alternative hypothesis is expressed as \neq ; whereas a one-tail hypothesis test involves the alternative hypothesis being stated as $<$ or $>$.

A Type I error occurs when the null hypothesis is rejected when, in reality, it is true. The probability of this error occurring is known as α , the

level of significance.

♦ A Type II error occurs when the null hypothesis is accepted when, in reality, it is not true. The probability of this error occurring is known as β .

(hapter

16

Hypothesis Testing with One Sample

In This Chapter

- ♦ Testing the mean of a population using a large and small sample
- ♦ Examining the role of alpha (α) in hypothesis testing
- ♦ Using the p-value to test a hypothesis
- ♦ Testing the proportion of a population using a large sample

In Chapter 15, I introduced the concept of hypothesis testing to whet your appetite. I have devoted this chapter to hypothesis testing that involves only one population, whereas in Chapter 17 I will discuss testing that compares two different populations to each other.

Hypothesis testing involving one population focuses on confirming claims such as the population average is equal to a specific value. We will consider many different cases with this type of hypothesis testing in the following sections. This chapter relies on many of the concepts we explored in Chapters 14 and 15, so be sure you are comfortable with that material before you dive into this chapter.

228 Part II: Inferential Statistics

Hypothesis Testing for the Mean with Large Samples

When the sample size we use to test our hypothesis is large ($n > 30$), we can rely on our old friend the central limit theorem which we met in Chapter 13. However, we still have two cases to consider—whether σ , the population standard deviation, is known or unknown.

When Sigma Is Known

To demonstrate this type of hypothesis test, I'll use the following story.

One of the most feared phrases a husband can hear from his wife is, "Honey, let's go on a diet together." I should have been suspicious of Debbie's motives when she suggested we go on the low-carbohydrate diet, especially because she wears size 2 pants. But I guess I could stand to lose a few pounds, so in a weak moment, I agreed. After all, I figured we could turn this into a competition to make things more interesting.

After a few harrowing days without my beloved carbohydrates (who would have guessed a grown man could dream about Cheez-its night after night), I began to wonder how Debbie was doing so well with the diet. I found the answer to this mystery hidden deep in the trunk of her car—a half-eaten box of cinnamon rolls. I guess that makes me the winner. The thrill of victory!

Anyway, let's say that this particular diet claims that the average age of the person who participates in this self-inflicted torture is less than 40 years old. We set up our hypothesis as follows:

$H_0 : \mu \geq 40$ years old

$H_a : \mu < 40$ years old

We sample 60 people on the diet and find that their average age is 35.7 years. Given that σ , the population standard deviation, is 16 years, we'll test the hypothesis at $\alpha = 0.05$.

Bob's Basics

Remember from Chapter 15 that α , the level of significance, represents the probability of making a Type I error. A Type I error occurs when we reject H_0 , when H_0 is true.

is actually true. In this case, a Type I error would mean that we believe the claim that the average person on the diet is less than 40 years old when, in reality, the claim is not true. For this example there's a 5 percent chance of this error happening.



Because the sample size is greater than 30 and we know the value of σ , we calculate the z-score from the standardized normal distribution as we did in Chapter 15.

x - IV

a-For our example, the standard error of the mean, $\sigma_{\bar{x}}$, would be:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.07}{\sqrt{16}} = 0.5175 \text{ years}$$

This results in a calculated z-score of:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{35.7 - 40}{0.5175} = -8.12$$

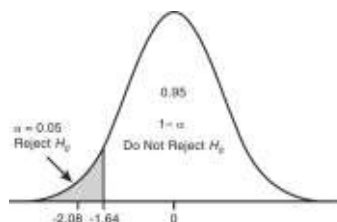
$$z = -8.12$$

$$z = -8.12$$

Also recall from Chapter 15, the critical z-score, which defines the boundary for the rejection region, is -1.64 for a one-tail (left side) test with $\alpha = 0.05$. Figure 16.1 shows this test graphically.

Figure 16.1

One-tail hypothesis test for the diet example (standardized scale).



$$\alpha = 0.05 \text{ Reject } H_0$$

Number of Standard Deviations from the Mean

As you can see in the figure, the calculated z-score of -2.08 falls within the "Reject H_0 " region, which allows us to conclude that the claim that the average age of those on this diet is less than 40 years old. I knew I was too old for this diet! In general, we reject H_0 if $|z| > z_{\alpha}$, where $|z|$ means the "absolute value of z." For instance, $|-2.08| = 2.08$.

When Sigma Is Unknown

Many times, we just don't have enough information to know the value of σ , the population standard deviation. However, as long as our sample size is 30 or more, we can substitute s , the sample standard deviation for σ . To illustrate this technique, let's use the following example.

230 PartJ: Inferential Statistics

I don't know about you, but it seems I spend too much time on the phone waiting on hold for a live customer service representative. Let's say a particular company has claimed that the average time a customer waits on hold is less than five minutes. We'll assume we do not know the value of σ . The following table represents the wait time in minutes for a random sample of 30 customers.

Wait Time in Minutes

6.2 3.8 1.3 5.4 4.7 4.4 4.6 5.0 6.6 8.3 3.2 2.7 4.0 7.3 3.6 4.9 0.5 2.9 2.5 5.6
5.5 4.7 6.5 7.1 4.4 5.2 6.1 7.4 4.8 2.9

Using Excel, we can determine that $\bar{x} = 4.74$ minutes and $s = 1.82$ minutes. At first glance, it appears the company's claim is valid. But let's put it through a hypothesis test with $\alpha = 0.02$ to be sure.

State the hypothesis as:

$H_0 : \mu \geq 5.0$ minutes

$H_1 : \mu < 5.0$ minutes

From Chapter 15, we know that the critical z-score for a one-tail (left side)

hypothesis test with $\alpha = 0.02$ is -2.05 .

As we did earlier in Chapter 14, we can approximate the standard error of the mean

by:

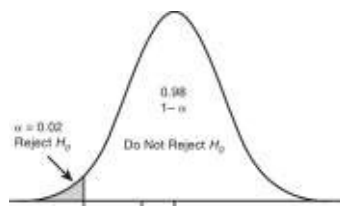
$$s = 1.82 \text{ minutes} \cdot \sqrt{30} = 0.332 \text{ minutes}$$

Our calculated z-score using this particular sample would be:

$$Z = \frac{4.74 - 5.0}{0.332} = -0.78$$

Figure 16.2 shows this test graphically.

According to our figure, we do not reject the null hypothesis. In other words, we do not have enough evidence from this sample to support the company's claim that the average wait on hold is less than five minutes. Even though the sample average is actually less than five minutes (4.74), it is too close to five minutes to say there is a difference between the two values. Another way to state this is to say: "The difference between 4.74 and 5.0 is not statistically significant in this case."



$$Z = -0.78$$

Number of Standard Deviations from the Mean

Figure 16.2

One-tail hypothesis test for waiting on hold example (standardized scale).

The Role of Alpha in Hypothesis Testing

For all the examples in these last two chapters, I have just stated a value for α , the level of significance. I'm sure you're wondering what impact

changing the value of α will have on the hypothesis test. Great question!

Suppose that I am making a claim that the average grade for a person using this book will be more than an 87. (I'm not really making this claim, so don't get too excited!) I would state the hypothesis test as follows:

ii,

$$H_0: \mu \leq 87 \quad H_a: \mu > 87$$

Now, it would be in my best interest if I could reject H_0 which would validate my claim. I can do so by choosing a fairly high value for α , say 0.10. This corresponds to a critical z-score of +1.28, because we are using the right tail of a one-tail hypothesis test.

Let's say that σ , the population standard deviation, is 12 and my sample mean is 90.6, which was taken from a sample size of 32 students. For this example, the standard error of the mean, $\sigma_{\bar{x}}$, would be:

$\sigma_{\bar{x}}$

$$\frac{\sigma}{\sqrt{n}}$$

$$\frac{12}{\sqrt{32}}$$

$$= 2.12$$

2J2 Part 3: Inferential Statistics

This results in a calculated z-score of: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{90.6 - 87}{2.12}$

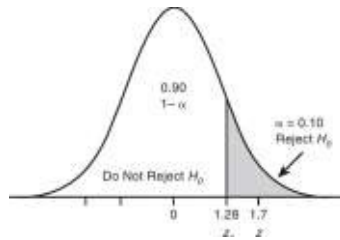
$$= \frac{3.6}{2.12}$$

$$= +1.70$$

According to Figure 16.3, I have achieved my goal of rejecting H_0 , because the calculated z-score is within the shaded region. My book appears to have done the trick!

Figure 16.3

Hypothesis test for gm example, $\alpha = 0.10$.

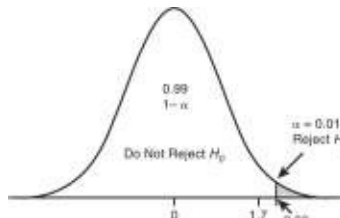


Number of Standard Deviations from the Mean

However, I must admit, I chose a pretty "wimpy" value of $\alpha = 0.10$ in an effort to help prove my claim. In this case, I am willing to accept a 10 percent chance of a Type I error. A more impressive test would be to set alpha lower, say $\alpha = 0.01$. Now that's a "real man's alpha." The level of significance corresponds to a critical z-score of $+2.33$. Figure 16.4 shows the impact of this change.

Figure 16.4

Hypothesis test for grade example, $\alpha = 0.01$.



2.33 z z c

Number of Standard Deviations from the Mean

Chapter 16: Hypothesis Testing with One Sample Z33

As you can see, to my horror, the shaded region no longer includes my calculated z-score of $+1.7$. Therefore, I do not reject H_0 and cannot claim the average grade of those using my book exceeds an 87. In general, a hypothesis test that rejects H_0 is most impressive with a low value of α .

Introducing the p-Value

def # i # rii # ti©n

Just when you thought it was safe to get back in the water, along comes another shark! This is the perfect opportunity to throw another concept at you. You might feel like grumbling a little right now, but in the end you'll be thanking me.

The p-value is the smallest level of significance at which the null hypothesis will be rejected, assuming the null hypothesis is true. The p-value is sometimes referred to as the observed level of significance. I know this may sound like a lot of mumbo-jumbo right now, but an illustration will help make this clear.

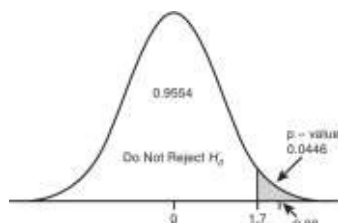
The observed level of significance is the smallest level of significance at which the null hypothesis will be rejected, assuming the null hypothesis is true. It is also known as the p-value.

The α -Value for a One-Tail Test

Using the previous grade example (over 87 if using this book), the p-value is represented by the shaded area to the right of the calculated z-score of +1.7. This is shown in Figure 16.5.

Figure 16.5

p-value for the gi-ade



Number of Standard Deviations from the Mean

Bob's Basics



Recall that $P[z > +1.7] = 1 - P[z < +1.7] = 1 - 0.9554 = 0.0446$. See

Chapter Vi if you need a refresher on using the standardized normal z table.



in

Using our standardized normal z table (Table 3 in Appendix B), we can confirm that the shaded area in the right tail is equal to $P[z > +1.7] = 0.0446$.

Because our p-value of 0.0446 is more than the value of α (set at 0.01), we do not reject H_0 . Most statistical software packages (including Excel) provide p-values with the analysis.

Bob's Basics

We can use the p-value to determine whether or not to reject the null hypothesis. In general ...

♦ If $p\text{-value} < \alpha$, we reject the null hypothesis.

If $p\text{-value} > \alpha$, we do not reject the null hypothesis.

Another way to describe this p-value is to say, in a very scholarly voice, "Our results are significant at the 0.0446 level." This means that as long as the value of α is 0.0446 or larger, we will reject H_0 , which is normally good news for researchers trying to validate their findings.

Calculating the p-value for a two-tail hypothesis test is slightly different, and I'll show you how in the next section.

The p -Value for a Two-Tail Test

Recall that you use a two-tail hypothesis test when the null hypothesis is stated as an equality. For example, let's test a claim that states the average number of miles driven by a passenger vehicle in a year equals 11,500 miles. I have serious reservations about this claim after spending half the day being a taxi driver to the kids. We would state the hypotheses as follows:

$H_0: \mu = 11,500$ miles

$H_a: \mu \neq 11,500$ miles

Let's assume $\sigma = 3000$ miles, and we want to set $\alpha = 0.05$. We sample 80 drivers and determine the average number of miles driven is 11,900. What is our p-value, and what do we conclude about the hypothesis?

For this example, the standard error of the mean, $\sigma_{\bar{x}}$, would be:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3000}{\sqrt{80}}$

$\sigma_{\bar{x}} = 335.41$ miles

This results in a calculated z-score of: $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{11,900 - 11,500}{335.41}$

$z = \frac{400}{335.41} = 1.19$

$\alpha = 0.05$

1.19

The critical z-score for a two-tail test with $\alpha = 0.05$ is ± 1.96 . The shaded area in Figure 16.6 shows the p-value for this test.

p-Value for a Two-Tail Hypothesis Test



Figure 16.6

p-value for the miles driven per year example.

According to Table 3 in Appendix B, the $P[z < +1.19] = 0.8830$. This means the shaded region in the right tail of Figure 16.6 is $P[z > +1.19] = 1 - 0.8830 = 0.117$. Because this is a two-tail test, we need to double this area to arrive at our p-value. According to our figure, the p-value is the total area of both shaded regions, which is $2 \times 0.117 = 0.234$. Because $p > \alpha$, we do not reject the null hypothesis. Our data supports the claim that the average number of miles driven per year by a passenger vehicle is 11,500.

In general, the smaller the p-value, the more confident we are about rejecting the null hypothesis. In most cases a researcher is attempting to find support for the alternative hypothesis. A low p-value provides support that brings joy to his or her heart.

236 Part 3: Inferential Statistics

Hypothesis Testing for the Mean with Small Samples

Recall, from Chapter 14, that with a small sample size, we lose the use of the central limit theorem, so, therefore, we need to assume that the population is normally distributed for all cases in this section. The first case that we'll examine is when we know σ , the population standard deviation.

When Sigma Is Known

When σ is known, the hypothesis test reverts back to the large sample size case. We can do this because we are now assuming the population is normally distributed. We can demonstrate this method with the following example.

Opening up my monthly cell phone bill lately has become a nerve-racking experience. As I warily open the envelope, I wonder what surprises await me. With several users on our family "share plan," I can often count on somebody having discovered a new feature that has nothing to do with talking to another person on the phone and having used this new-found discovery over and over and over again. Occasionally, after digging through countless pages full of numbers and codes, I breathe a sigh of relief and say a silent prayer of thanks. Most months, however, I end up clutching my chest and screaming "AIEEEEEEEEE!" It's like playing a subtle form of Russian roulette with the phone company.

Anyway, let's say the phone company claims that the average monthly cell phone bill for their customers is \$92 (I wish). We can test this claim by stating our hypothesis as:

$$H_0: \mu = \$92 \quad H_1: \mu \neq \$92$$

Bob's Basics



Recall from Chapter 14 that because we know σ and we assumed the population is normally distributed, we can use the z-scores from the normal probability distribution to test this hypothesis.

We'll assume that $\mu = \$22.50$ and that the population is normally distributed. We select 18 phone bills randomly and determine the sample average equals \$107. Using $\alpha = 0.02$, what do we conclude?

For this example, the standard error of the mean, $\sigma_{\bar{x}}$, would be:

Or

$$= \frac{\$22.50}{\sqrt{18}}$$

$$= \$5.30$$

This results in a calculated z-score of:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\$107 - \$92}{\$5.30}$$

$$z = \frac{\$15}{\$5.30} = +2.83$$

$$\alpha = 0.02$$

The critical z-score for a two-tail test with $\alpha = 0.02$ is ± 2.33 . Figure 16.7 shows this test graphically.

As you can see in Figure 16.7, the calculated z-score of $+2.83$ is within the "Reject H_0 " region. We, therefore, conclude that the average cell phone bill is not equal to \$92. I didn't think so!

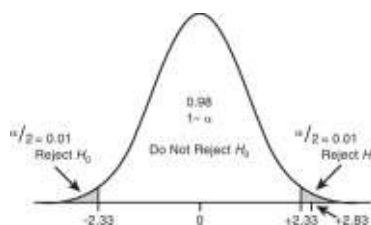


Figure 16.7

Hypothesis test for cell phone bills.

Number of Standard Deviations from the Mean

When Sigma Is Unknown

As we did in Chapter 14, when σ is unknown for a small sample size taken from a normally distributed population, we use the Student's t -distribution. This particular distribution allows us to substitute s , the sample standard deviation for σ .

As an example, suppose my son John claims his average golf score is less than 88. Not to be one to doubt him, I can test this claim with the following hypothesis:

$$H_0: \mu \geq 88$$

$$H_a: \mu < 88$$

We will assume that we do not know σ and that John's scores follow a normal distribution. The following represents a random sample of 10 golf scores from John.

Using Excel, we can determine that $\bar{x} = 86.3$ and $s = 2.58$ for this sample. Recall from Chapter 14, we can approximate the standard error of the mean using the following equation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{2.58}{\sqrt{10}} = 0.816$$

We can then determine the calculated t -score using the following equation: $\frac{\bar{x} - \mu_0}{s_{\bar{x}}}$, $\frac{86.3 - 88}{0.816}$

$$= -2.08$$

$$-2.08$$

We'll test this hypothesis using $\alpha = 0.05$. To find the corresponding

critical t-score, we use Table 4 from Appendix B. Here is an excerpt of this table.

Student's t-Distribution Table

You recall from Chapter 14, we need to determine the number of degrees of freedom, which is equal to $n - 1 = 10 - 1 = 9$ for this example. Because this is a one-tail (left side) test, we look under the $\alpha = 0.05$ column resulting in a critical t-score, t , equal to -1.833, which is underlined. Figure 16.8 shows this test graphically.

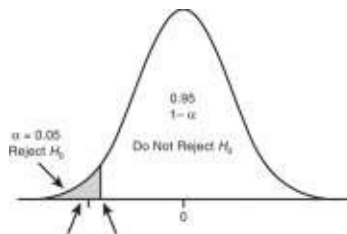


Figure 16.8

Hypothesis test for John's golf

As we can see in the figure, the calculated t-score of -2.08 falls within the shaded "Reject H_0 " region. Therefore, we can conclude that John's average golf score is indeed lower than 88. So that explains why he usually beats me! In general, we reject H_0 if $|t| > |t_r|$.

Let's take another example to demonstrate a two-tail hypothesis test using the t-distribution. I would like to test a claim that the average speed of cars passing a specific spot on the interstate is 65 miles per hour. We can express the hypothesis test as follows:

$H_0 : \mu = 65 \text{ miles per hour}$

$H_a : \mu \neq 65 \text{ miles per hour}$

We will assume that we do not know σ and that speeds follow a normal distribution. The following represents a random sample of the speed of seven cars.

Bob's Basics



Because John's golf score example is a one-tail test on the left side of the distribution, we use a negative critical t-score. Had this been a one-tail test on the right side, we would use a positive critical t-score.

Bob's Basics

It is not possible to determine the p-value for a hypothesis test when using the Student's t-distribution table in Appendix B. However, most statistical software will provide the p-value as part of the standard analysis. We'll see this in later chapters as we use Excel.

Part 3: Inferential Statistics

Car Speeds

62 74 65 68 71 64 68

Using Excel, we can determine that $\bar{x} = 66.9$ mph and $s = 4.16$ mph for this sample. We can approximate the standard error of the mean:

$$s / \sqrt{n} = 4.16 / \sqrt{7} = 1.57 \text{ mph}$$

We can then determine the calculated t-score: $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{66.9 - 65}{1.57} = 1.21$

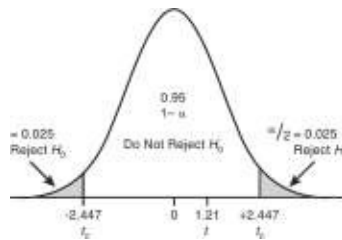
We'll test this hypothesis using $\alpha = 0.05$. To find the corresponding critical t-score, we use Table 4 from Appendix B. Here is an excerpt of this table.

Student's t-Distribution Table

The number of degrees of freedom for this example equals $n - 1 = 7 - 1 = 6$. Because this is a two-tail test, we need to divide $\alpha = 0.05$ into two equal portions, one on the right side of the distribution, the other on the left. We then look under the $\alpha/2 = 0.025$ column resulting in a critical t-score, $t_{\alpha/2}$, equal to + 2.447, which is underlined. This test is shown graphically in Figure 16.9.

Chapter 16: Hypothesis Testing with One Sample z

Hypothesis Test for Car Speeds (Two Tail t -Distribution)



$\alpha/2 = 0.025$ Reject H_0

1.21 +2.447

Figure 16.9

Hypothesis test for car speed

As we can see in the figure, the calculated t -score of +1.21 falls within the "Do Not Reject H_0 " region. Therefore, we can conclude that the average speed past this spot on the interstate averages 65 miles per hour.

Using Excel's TINV Function

We can generate critical t -scores using Excel's TINV function, which has the following characteristics:

TINV(probability, deg-freedom)

where:

probability = the level of significance, α , for a two-tail test

deg-freedom = the number of degrees of freedom

For instance, Figure 16.10 shows the TINV function being used to determine the critical t -score for $\alpha = 0.05$ and d.f. = 6 from our previous example, which was a two-tail test.

Figure 16.10

Excel's TINV function for t two-tail test.

2tt Part 3: Inferential Statistics

Cell A1 contains the Excel formula =TINV(0.05, 6) with the result being 2.447. This probability is underlined in the previous table.

A one-tail test requires a slight modification. We need to multiply the probability in the TINV function by two because this parameter is based on a two-tail test. Figure 16.11 shows the TINV function being used to determine the critical t-score for $\alpha = 0.05$ and d.f. = 9 from our earlier one-tail test example with John's golf scores.

Figure 16.11

Excel's TINV function for t one-tail test.

Cell A1 contains the Excel formula =TINV(2*0.05, 9) with the result being 1.833. This is consistent with the result from our previous example.

Hypothesis Testing for the Proportion with Large Samples

You can perform hypothesis testing for the proportion of a population as long as the sample size is large enough. Recall from Chapter 13, that proportion data follows the binomial distribution, which can be approximated by the normal distribution under the following conditions:

$$np > 5 \text{ and } nq > 5$$

where:

p = the probability of a success in the population
 q = the probability of a failure in the population ($q = 1 - p$)

We will examine both one-tail and two-tail hypothesis testing for the proportion in the following sections.

Chapter 16: Hypothesis Testing with One Sample z and t

One-Tail Hypothesis Test for the Proportion

Let's say we would like to test the hypothesis that more than 30 percent of U.S. households have Internet access. We would state the hypothesis as:

$$H_a: p > 0.30$$

$$H_0: p \leq 0.30$$

where p = the proportion of U.S. households with Internet access.

We collect a sample of 150 households and find that 38 percent of these have Internet access. What can we conclude at the $\alpha = 0.05$ level?

Our first step is to calculate σ_p , the standard error of the proportion, which was described in Chapter 13 using the following equation:

$$\sigma_p = \sqrt{p(1-p)}$$

, Wrong Number

$\sigma_p = \sqrt{p(1-p)}$ Be careful not to confuse

this definition of p with the p -value that we talked about earlier.

$$\sigma_p = \sqrt{p(1-p)}$$

$$\sigma_p = \sqrt{(0.30)(1-0.30)}$$

where p_H = the proportion assumed by the null hypothesis. For our example:

$$\sigma_p = 0.037$$

Next, we can determine the calculated z -score using:

where p = the sample proportion. For our example: $P - p_H = 0.38 - 0.30$

$$0.037$$

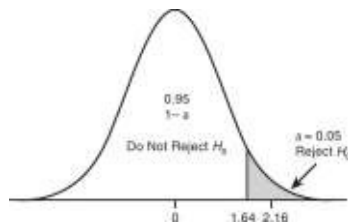
= +2.16

The critical z-score for a one-tail test with $\alpha = 0.05$ is +1.64. This hypothesis test is shown graphically in Figure 16.12.

Ikk Part 3: Inferential Statistics

Figure 16.12

Hypothesis test for the Internet access example.



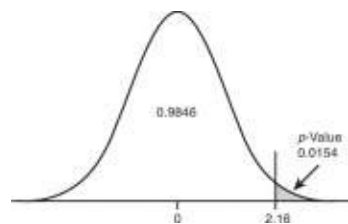
Number of Standard Deviations from the Mean

As you can see in Figure 16.12, the calculated z-score of +2.16 is within the "Reject H_0 " region. Therefore, we conclude that the proportion of U.S. households with Internet access exceeds 30 percent.

We can show the p-value for this test graphically in Figure 16.13.

Figure 16.13

p-value for the Internet access example.



p-Value 0.0154

Number of Standard Deviations from the Mean

Using our standardized normal z table (Table 3 in Appendix B), we can confirm that the shaded area in the right tail is equal to:

$$P[z > +2.16] = 1 - P[z < +2.16]$$

$P[z > +2.16] = 1 - 0.9846 = 0.0154$ Therefore, our results are significant at the 0.0154 level. As long as $\alpha > 0.0154$, we will be able to reject H_0 .

Chapter 16: Hypothesis Testing with One Sample ZW

Two-Tail Hypothesis Test for the Proportion

We'll wrap this chapter up with one final two-tail example. Here, we want to test a hypothesis for a company that claims 50 percent of their customers are of the male gender. We state our hypothesis as:

$$H_0: p = 0.50$$

$$H_a: p \neq 0.50$$

We randomly select 256 customers and find that 47 percent are male. What can we conclude at the $\alpha = 0.05$ level?

We need to determine σ_p , the standard error of the proportion:

$$\sigma_p = \sqrt{p_0(1-p_0)} = \sqrt{(0.50)(1-0.50)}$$

$$= 0.50$$

Next, we can determine the calculated z-score: $Z = \frac{p - p_0}{\sigma_p} = \frac{0.47 - 0.50}{0.50}$

$$= -0.0312$$

$$= -0.0312$$

$$= -0.96$$

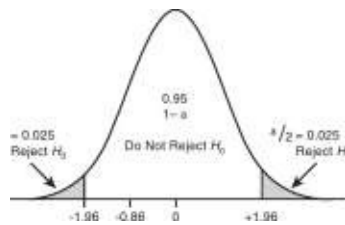
The critical z-score for a two-tail test with $\alpha = 0.05$ is ± 1.96 . This hypothesis test is shown graphically in Figure 16.14.

As you can see in Figure 16.14, the calculated z-score of -0.96 is within the "Do Not Reject H_0 " region. There, we conclude that the proportion of male customers is equal to 50 percent for this company.



Bob's Basics

In general, we reject H_0 if $|z| > |z_c|$ or $|f| > |f_c|$. Also, we do not reject H_0 if $|z| < |z_c|$ or $|f| < |f_c|$.



$$\alpha / 2 = 0.025$$

Reject H_0 if

$$|z| > 1.96$$

Reject H_0 if

Figure 16.15

Hypothesis test for the percentage of males example.

Number of Standard Deviations from the Mean

2.16 Part J: Inferential Statistics

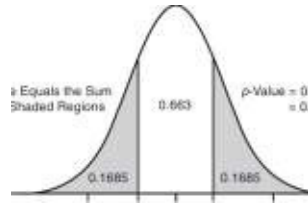
Figure 16.15 graphically shows the p-value for this test.

Figure 16.15

p-value for the percentage of

males example.

p-Value Equals the Sum of the Shaded Regions



$$p\text{-Value} = 0.1685 + 0.1685 = 0.337$$

~i 1 r

$$-0.860 + 0.86 + 1.96$$

Using our standardized normal z table (Table 3 in Appendix B), we can confirm that the shaded area in the left tail is equal to:

$$P[z < -0.86] = 1 - P[z < +0.86]$$

$P[z < -0.86] = 1 - 0.8315 = 0.1685$ Because this is a two-tail test, the *i*-value would be $2 \times 0.1685 = 0.337$, which represents the total area in both shaded regions.

Your Turn

1. Test the claim that the average SAT score for graduating high school students is equal to 1100. A random sample of 70 students was selected, and the average SAT score was 1035. Assume $\sigma = 310$ and use $\alpha = 0.10$. What is the *R*-value for this sample?

2. A student organization at a small business college claims that the average class size is greater than 35 students. Test this claim at $\alpha = 0.02$, using the following sample of class size:

42 28 36 47 35 41 33 30 39 48

Assume the population is normally distributed and that σ is unknown.

3. Test the claim that the average gasoline consumption per car in the United States is more than 7 liters per day. (We're going metric here!) Use the random sample here, which represents daily gasoline usage for one car:

9 6 4 12 4 3 18 10 4 5

3 8 4 11 3 5 8 4 12 10

9 5 15 17 6 13 7 8 14 9

Assume the population is normally distributed, and that σ is unknown. Use $\alpha = 0.05$ and determine the p-value for this sample.

4. Test the claim that the proportion of Republican voters in a particular city is less than 40 percent. A random sample of 175 voters was selected and found to consist of 30 percent Republicans. Use $\alpha = 0.01$ and determine the p-value for this sample.

5. Test the claim that the proportion of teenage cell phone users exceeding their allotted monthly minutes equals 65 percent. A random sample of 225 teenagers was selected and found to consist of 69 percent exceeding their minutes. Use

$\alpha = 0.05$ and determine the p-value for this sample.

6. Test the claim that the mean number of hours that undergraduate students work at a particular college is less than 15 hours per week. A random sample of 60 students was selected, and the average number of working hours was 13.5 hours per week. Assume $\sigma = 5$ hours, and use $\alpha = 0.10$. What is the p-value for this sample?

The Least You Need to Know

◆ The smaller the value of α , the level of significance; the more difficult it is to reject the null hypothesis.

◆ We reject H_0 if $|z| > |z_t|$ or $|t| > |t_t|$.

◆ The p-value is the smallest level of significance at which the null hypothesis will be rejected, assuming the null hypothesis is true.

◆ If the p-value $< \alpha$, we reject the null hypothesis. If p-value $> \alpha$, we do not reject the null hypothesis.

◆ Use the Student's t-distribution for the hypothesis test when $n < 30$, σ is unknown, and the population is normally distributed.

(hapte

17

Hypothesis Testing with Two Samples

In This Chapter

<?> Developing the sampling distribution for the difference in means

◆ Testing the difference in means between populations using a large and small sample

◆ Distinguishing between independent and dependent samples

◆ Using Excel to perform a hypothesis test

◆ Testing the difference in proportions between populations

Now we're really cooking. Because you have done so well with one sample hypothesis testing, you are ready to graduate to the next level—two-sample testing. Here we often test to see whether there is a difference between two separate populations. For instance, I could test to see whether there was a difference between Brian and John's average golf score. But being an "experienced" parent, I know better than to go near that one.

Because many similarities exist between the concepts of this chapter and those of Chapter 16, you should have a firm handle on the previous chapter's material before you jump into this one.

250 Part 3: Inferential Statistics

The Concept of Testing Two Populations

Many statistical studies involve comparing the same parameter, such as a mean, between two different populations. For example:

- ♦ Is there a difference in average SAT scores between males and females?
- ♦ Do "long-life" light bulbs really outlast standard light bulbs?

def'i'rrtienn

The sampling distribution for the difference in means

describes the probability of observing various intervals for the difference between two sample means.

- ♦ Does the average selling price of a house in Newark differ from the average selling price for a house in Wilmington?

To answer such questions, we need to explore a new sampling distribution. (I promise this will be the last.) This one has the fanciest name of them all—the sampling distribution for the difference in means. (Dramatic background music brings us to the edge of our seats.)

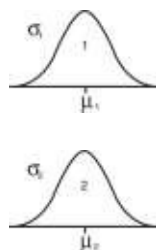
Sampling Distribution for the Difference in Means

The sampling distribution for the difference in means can best be described in Figure 17.1.

Figure 17.1

The sampling distribution for the difference in means.

Population 1



Sampling Distribution for the Mean (Population 1)



$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

M-1

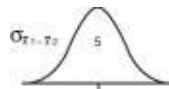
Sampling Distribution for the Mean (Population 2)

G:



Hx2

Sampling Distribution for the Difference in Means



Hx,-x

As an example, let's consider testing for a difference in SAT scores for male and female students. We'll assign female students as Population 1 and male students as Population 2. Graph 1 in Figure 17.1 represents the distribution of SAT scores for the female students with mean μ_1 , and standard deviation σ_1 . Graph 2 represents the same for the male population.

Graph 3 represents the sampling distribution for the mean for the female students. This graph is the result of taking samples of size n , and plotting the distribution of sample means. Recall that we discussed this distribution of sample means back in Chapter 13. The mean of this distribution would be:

$$\mu_{\bar{x}} = \mu_1$$

This is according to the central limit theorem from Chapter 13. The same logic holds true for Graph 4 for the male population.

Graph 5 in Figure 17.1 shows the distribution that represents the

difference of sample means from the female and male populations. This is the sampling distribution for the difference in means, which has the following mean:

$\mu_{\bar{X} - \bar{Y}}$

$= \mu_X - \mu_Y$,

In other words, the mean of this distribution, shown in Graph 5, is the difference between the means of Graphs 3 and 4.

The standard deviation for the Graph 5 is known as the standard error of the difference between two means and is calculated with:

$\sigma_{\bar{X} - \bar{Y}}$

$= \sqrt{\sigma_X^2 + \sigma_Y^2}$;

where:

σ_X^2 and σ_Y^2

are the variances for Populations 1 and 2

n_1 and n_2 = the sample size from Populations 1 and 2

Now before you pull the rest of your hair out, let's put these guys to work in the following section.

Definition

• **Standard Error of the Difference Between Two Means**

The standard error of the difference between two means

describes the variation in the difference between two sample means and is calculated using:

$\sigma_{\bar{X} - \bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$

where n_1 and n_2

Testing for Differences Between Means with Large Sample Sizes

When the sample sizes from both populations of interest are greater than 30, the central limit theorem allows us to use the normal distribution to approximate the sampling distribution for the difference in means. Let's demonstrate this technique with the following example.

Studies have been done to investigate the effects of stimulation on the brain development of rats. I guess the logic being what's good for rats can't be all that bad for us humans. Two samples were randomly selected from the same rat population.

The first sample, we'll call these the "lucky rats" (Population 1), was surrounded with every luxury a rat could imagine. I can envision a country club atmosphere, complete with a golf course (and tiny golf carts), tennis courts, and a five-star restaurant where our lucky rats could feast on imported cheese and French wine while they discussed the state of the rat economy.

The second sample, we'll call them the "less-fortunate rats" (Population 2), didn't have it quite so good. These guys were locked in a barren cage and were forced to eat Cheez Whiz from a can and watch reruns of reality TV shows. Animal rights activists protested against this experiment, claiming the involuntary use of Cheez Whiz was "inhumane."

After spending three months in each of these environments, the size of each rat brain was measured by weight for development. I'll spare you the details as to how this was done, but I will tell you that Harvey the Rat mysteriously failed to show for his 8 A.M. tee time. His group went off without him.

The following table summarizes these gruesome findings.

Summarized Data for Rat Experiment

Average Brain Sample Standard Sample

Population Weight in Grams Deviation Size

x s n

Lucky (1) 2.4 0.6 50

Less-Fortunate (2) 2.1 0.8 60

For this hypothesis test, we need to assume that the two samples are independent of each other. In other words, there is no relationship between the rats in the lucky

sample and the rats in the less-fortunate sample. The hypothesis statement for this two-sample test would be as follows:

$H_0: \mu_1 \leq \mu_2$,

$H_a: \mu_1 > \mu_2$,

where:

μ_1 = the mean brain weight of the lucky rat population

μ_2 = the mean brain weight of the less-fortunate rat population

The hypothesis can also be expressed as:

$\mu_1 - \mu_2 < 0$

$H_a: \mu_1 - \mu_2 > 0$

The alternative hypothesis supports the claim that the lucky rats will have heavier brains. Seems to me this could lead to neck problems for these rats—but I'll leave that question for another study. We'll test this hypothesis at the $\alpha = 0.05$ level.

If σ_1 or σ_2 are not known, then we can use s_1 or s_2 , the standard deviation from the samples of populations 1 and 2 as an approximation, as long as $n > 30$ for both populations, as shown here:

With this assumption, we can approximate the standard error of the difference between two means using:

ft] n 2

Because we do not know σ , or σ , in our rat example, we set: $d, = s$, and

$la; a; l(0.6)^2 (0.8)^2 <>: : = J- + - = \backslash rwm = 0A34 \text{ gramS}$

25<i Part 3: Inferential Statistics

Bob's Basics



n

We are now ready to determine the calculated z-score using the following equation:

$$(x_1 - x_2) - (u_1 - u_2)$$

$$z = -$$

$$a - -$$

For the rat example, our calculated z-score becomes:

$$fc - ^{-} - Qi, - ti :), o _ (2.4 - 2.1) - o _$$

$$c - _ 0.134$$

Figure 17.2 shows the results of this hypothesis test.

The critical z-score for a one-tail (right side) test with $\alpha = 0.05$ is +1.64. According to Figure 17.2, this places the calculated z-score of +2.24 in the "Reject

H" region, which leads to our conclusion that the lucky rats have heavier brains than

the less-fortunate rats.

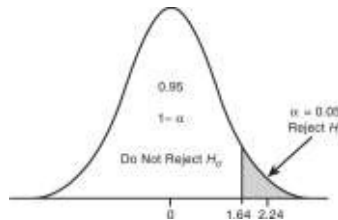
The term $(\mu_1 - \mu_2)$ refers to the hypothesized difference between the

two population means. When the null hypothesis is testing that there is no difference between population means, then the term

$(n, -H_2)$, is set to 0.

Figure 17.2

Hypothesis test for rat example.



Number of Standard Deviations from the Mean

Random Thoughts

The conditions that are necessary for the hypothesis test for differences between means with large sample sizes are as follows:

- ◆ The samples are independent of each other.
- ◆ The size of each sample must be greater than or equal to 30.
- ◆ If the population standard deviations are unknown, we can use the sample standard deviations to approximate them.

We can find the p-value for this sample by using the normal z-score table found in Appendix B as follows:

$$P[z > +2.24] = 1 - P[z < +2.24] \quad P[z > +2.24] = 1 - 0.9875 = 0.0125$$

Bob's Basics



We can also apply this technique to hypothesis tests that involve sample sizes less than 30. However, to do so, the following conditions must be

met:

- ◆ Both populations must be normally distributed.
- ◆ Both population standard deviations must be known.

The results of our rat study can greatly improve the lives of many. When your spouse catches you sneaking off to the golf course on Saturday morning, you can tell him or her with a straight face that you are just trying to improve your mind. We now have the statistics to support you. But be warned, you might develop a sore neck with all that extra brain weight.

Testing a Difference Other Than Zero

In the previous example, we were just testing whether or not there was any difference between the two populations. We can also test whether the difference exceeds a certain value. As an example, suppose we want to test the hypothesis that the average salary of a mathematician in New Jersey exceeds the average salary in Virginia by more than \$5,000. We would state the hypotheses as follows:

$$H_0: \mu_1 - \mu_2 \leq 5000 \quad H_a: \mu_1 - \mu_2 > 5000$$

where:

\bar{X}_1 = the mean salary of a mathematician in New Jersey | \bar{X}_2 = the mean salary of a mathematician in Virginia

We'll assume that $\sigma_1 = \$8100$ and $\sigma_2 = \$7600$, and we'll test this hypothesis at the $\alpha = 0.10$ level.

256 Part 3: Inferential Statistics

A sample of 42 mathematicians from New Jersey had a mean salary of \$51,500, whereas a sample of 54 mathematicians from Virginia had a mean salary of \$45,400.

The standard error of the difference between two means is:

a, a;

8100)" (7600)"

42

54

= \$1622.3

Our calculated z-score becomes:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\$51,500 - \$45,400) - (\$5,000)}{\sqrt{\frac{1622.3^2}{42} + \frac{1622.3^2}{54}}}$$

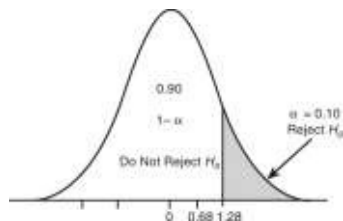
1622.3

: +0.68

The results of this hypothesis test are shown in Figure 17.3.

Figure 17.5

Hypothesis test for the salary example.



z z c Number of Standard Deviations from the Mean

The critical z-score for a one-tail (right side) test with $\alpha = 0.10$ is +1.28. According to Figure 17.3, this places the calculated z-score of +0.68 in the "Do Not Reject H" region, which leads to our conclusion that the difference in salaries between the two states does not exceed \$5,000.

Testing for Differences Between Means with Small Sample Sizes and Unknown Sigma

This section addresses the situation where the population standard

deviation, σ , is not known and the sample sizes are small. If one or both of our sample sizes is less than 30, the population needs to be normally distributed to use any of the following techniques. We made the same assumption for small sample sizes back in Chapters 14 and 16.

The sampling distribution for the difference between sample means for this scenario follows the Student's t -distribution. Also for small sample sizes, the equation for the standard error of the difference between two means, $O - \bar{O}$, depends on whether or not the standard deviations (or the variances) of the two populations are equal. The first example will deal with equal standard deviations.

Equal Population Standard Deviations

We have a very mysterious occurrence in our household—batteries seem to vanish into thin air. So I started buying them in 24-packs at the warehouse store, naively thinking that "these will last a long time." Wrong again—the more I buy, the faster they disappear. Maybe it has something to do with certain teenagers listening to music on their portable CD players at a "brain-numbing" volume into the wee hours of the morning. Just a thought. So if I ever hear about a new "longer-lasting battery," I'm all over it. Let's say a company is promoting one of these batteries, claiming that its life is significantly longer than regular batteries. The hypothesis statement would be:

$$H_0 : \mu_1 \leq \mu_2,$$

$$H_1 : \mu_1 > \mu_2,$$

where:

μ_1 = the mean life of the long-lasting batteries | μ_2 = the mean life of the regular batteries

We'll test this hypothesis at the $\alpha = 0.01$ level. The following data was collected measuring the battery life in hours for both types of batteries:

Raw Data for Battery Example

Long-Lasting Battery (Population 1):

51 44 58 36 48 53 57 40 49 44 60 50

Regular Battery (Population 2):

42 29 51 38 39 44 35 40 48 45

Using Excel, we can summarize this data in the following table.

258 Part 3: Inferential Statistics

In this example, we are assuming that $\sigma_1 = \sigma_2$, but that the values of μ_1 and μ_2 are unknown. Under these conditions, we calculate a pooled estimate of the standard deviation using the following equation:

$$l(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$$

$$n_1 + n_2 - 2$$

Don't panic just yet. This equation looks a whole lot better with numbers plugged in.

$$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 = (12 - 1)(7.31)^2 + (10 - 1)(6.40)^2$$

$$n_1 + n_2 - 2 = 12 + 10 - 2$$

$$s_p^2 = \frac{11(54.02) + 9(40.96)}{20} = 6.02$$

We can now approximate the standard error of the difference between two means using:

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Let's apply our example to this fellow. $= (6.92)$

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\alpha = 0.01$ We are now ready to determine our calculated t-score using the following equation: $t = \frac{(\bar{m}_1 - \bar{m}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$

2.96

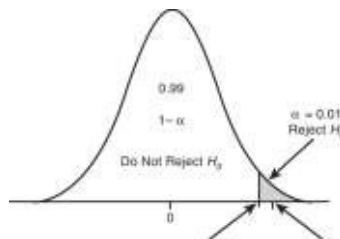
= +2.73

The number of degrees of freedom for this test are: $d.f. = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$

The critical t-score, taken from Table 4 in Appendix B, for a one-tail (right) test using $\alpha = 0.01$ with $d.f. = 20$ is +2.528. This hypothesis test is shown graphically in Figure 17.4.

Figure 17.A

Hypothesis test for the battery example.



According to Figure 17.4, our calculated t-score of +2.73 is found in the "Reject H_0 region, which leads to our conclusion that the long-lasting batteries do indeed have a longer life than the regular batteries. Now that has my attention!

This procedure was based on the assumption that the standard deviations of the populations were equal. What if this assumption is not true? Fm glad you asked!

260 PartJ: Inferential Statistics

Random Thoughts

The conditions that are necessary for the hypothesis test for differences between means with small sample sizes are as follows:

- ◆ The samples are independent of each other.
- ◆ The population must be normally distributed.
- ◆ If σ_1 and σ_2 are known, use the normal distribution to determine the rejection region.
- ◆ If σ_1 and σ_2 are unknown, approximate them with s_1 and s_2 and use the Student's t-distribution to determine the rejection region.

Unequal Population Standard Deviations

We'll investigate this scenario using the same battery example, but now we will assume that $\sigma_1 \neq \sigma_2$. The procedure is identical to the previous method except for two changes.

The first difference involves the standard error of the difference between two means. The equation used for this scenario is as follows:

$$s_p = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For the battery example, our result is:

$$s_p = \sqrt{\frac{(7.31)^2}{12} + \frac{(6.40)^2}{10}}$$

$$s_p = 2.92$$

$$s_p = 2.92$$

$$s_p = \sqrt{(4.45)^2 + (4.10)^2} = 2.92$$

We are now ready to determine our calculated t-score using the following equation: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$t = \frac{49.2 - 48.0}{2.92 \sqrt{\frac{1}{12} + \frac{1}{10}}}$$

$$t = +2.77$$

The second difference (hold on to your hat) is the method for determining the number of degrees of freedom for the Student's t-distribution.

d.f.

72, -1 72, -1

Before you have a seizure, let me demonstrate that this animal's bark is worse than its bite. First, recognize that for our battery example:

$\hat{\mu}_1 = 4.45$ and $\hat{\mu}_2 = 4.10$

12

10

We can now plug these values into the above equation as follows:

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2) \pm t_{\alpha} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\frac{(4.45 - 4.10) \pm 1.992 \sqrt{1.80 + 1.87}}{\sqrt{1.80 + 1.87}}$$

$$\frac{0.35 \pm 1.992 \sqrt{3.67}}{\sqrt{3.67}}$$

Because the number of degrees of freedom must be an integer, we round this result to 20. The critical t-score, taken from Table 4 in Appendix B, for a one-tail (right) test using $\alpha = 0.01$ with d.f. = 20 is +2.528. Because $t > t_c$, we reject H_0 .

Letting Excel Do the Grunt Work

Excel performs many of the hypothesis tests that we've discussed in this chapter. So let me explain how to perform the previous battery example using this nifty tool. Follow these steps:

1. Open a blank Excel sheet and enter the data from the battery example in Columns A and B as shown in Figure 17.5.
2. From the Tools menu, choose Data Analysis and select t-Test: Two-Sample Assuming Unequal Variances. (Refer to the section "Installing the Data Analysis Add-in" from Chapter 2 if you don't see the Data Analysis command on the Tools menu.)

Figure 17.5

Data entry for the battery example.

262 Part 3: Inferential Statistics

3. Click OK.

4. In the t-Test: Two-Sample Assuming Unequal Variances dialog box, choose cells B1:B12 for Variable 1 Range and cells A1:A10 for Variable 2 Range. Set the Hypothesized Mean Difference to 0, Alpha to 0.01, and Output Range to cell D1, as shown in Figure 17.6.

Figure 17.6

The t-test: Two-Sample Assuming Unequal Variances dialog box.



5. Click OK. The t-test output is shown in Figure 17.7.

Figure 17.7

t-test output.

According to Figure 17.7, the calculated t-score of 2.758 is found in cell E9, which differs slightly from what we calculated in the previous section (2.77) due to the rounding of numbers. The Rvalue of 0.006 is found in cell E10. Because $p\text{-value} < \alpha$, we reject the null hypothesis.

Testing for Differences Between Means with Dependent Samples

Up to this point, all the samples that we have used in the chapter have been independent samples. Samples are independent if they are not related in any way with each other. This is in contrast to dependent samples, where each observation of one sample is related to an observation in another.

def'i'mi'ticin

An example of a dependent sample would be a weight-loss study. Each person is weighed at the beginning (Population 1) and end (Population 2) of the program. The change in weight of each person is calculated by subtracting the Population 2 weights from the Population 1 weights. Each observation from Population 1 is matched to an observation in Population 2. Dependent samples are tested differently than independent samples.

To demonstrate testing dependent samples, let's revisit my golf ball example from Chapter 15. If you remember, I dreamed I had invented a golf ball that I claimed would increase the distance off the tee by more than 20 yards. To test my claim, suppose we had nine golfers hit my golf ball and the same golfers hit a regular golf ball. The following table shows these results. The letter "d" refers to the difference between my ball and the other ball.

With independent samples,

there is no relationship in the observations between the samples. With dependent samples,

the observation from one sample is related to an observation from another sample.

Distance in Yards for Golf Ball Example

For future calculations, we will need:

$$\sum d = 14 + 15 + 26 + 31 + 30 + 29 + 27 + 23 + 31 = 226$$

$$\sum V = 196 + 225 + 676 + 961 + 900 + 841 + 729 + 529 + 961 = 6018$$

16k PartJ: Inferential Statistics

The distances using my golf ball will be considered Population 1, and the distances with the other golf ball will be labeled Population 2. Because the same golfer hit both balls in each instance in the preceding table, these two samples are considered dependent.

My hypothesis statement for my claim would look like:

$$H_0: \mu_d \leq 20$$

$$H_a: \mu_d > 20$$

where:

\bar{X}_1 = the average distance off the tee with my new golf ball | \bar{X}_2 = the average distance off the tee with the other golf ball

However, because we are only interested in the difference between the two populations, we can rewrite this statement as a single sample hypothesis as follows:

$$H_0: \mu_d \leq 20 \quad H_a: \mu_d > 20$$

where μ_d is the mean of the difference between the two populations.

We will test this hypothesis using $\alpha = 0.05$.

Our next step is to calculate the mean difference, \bar{d} , and the standard deviation of the difference, s_d , between the two samples as follows:

$$\bar{d} = \frac{\sum d}{n} = \frac{226}{25} = 9.04$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{2260 - \frac{(226)^2}{25}}{25-1}}$$

$$= \sqrt{\frac{2260 - 2029.04}{24}}$$

$$= \sqrt{9.58}$$

$$= 3.09 \text{ yards}$$

The equation for s_d is the same standard deviation equation that you learned in Chapter 5.

If both populations follow the normal distribution, we use the Student's t-distribution because both sample sizes are less than 30 and σ_d and σ are unknown. The calculated t-score is found using:

$$l = \frac{25.11 - 20}{5.11}$$

$$t = \frac{l - \bar{l}}{s_l} = \frac{2.34}{1.0} = +2.34$$

$$J_i = 6 \text{ SI } 2.18$$

The number of degrees of freedom for this test is: $d.f. = n - 1 = 9 - 1 = 8$

The critical t-score, taken from Table 4 in Appendix B, for a one-tail (right) test using $\alpha = 0.05$ with $d.f. = 8$ is $+1.86$. This hypothesis test is shown graphically in Figure 17.8.

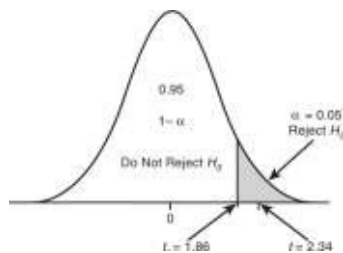


Figure 17.8

Hypothesis test for the golf ball example.

According to Figure 17.8, our calculated t-score of $+2.34$ is found in the "Reject H_0 " region, which leads to our conclusion that my golf ball increases the distance off the tee by more than 20 yards. Too bad this was only a dream!

Testing for Differences Between Proportions with Independent Samples

We can perform hypothesis testing to examine the difference between proportions of two populations as long as the sample size is large enough. Recall from Chapter 13, proportion data follow the binomial distribution, which can be approximated by the normal distribution under the following conditions.

266 Part 3: Inferential Statistics

$$np > 5 \text{ and } nq > 5$$

where:

p = the probability of a success in the population

q = the probability of a failure in the population ($q = 1 - p$)

Let's say that I want to test the claim that the proportion of males and females between the ages of 13 and 19 who use instant messages (IM) on the Internet every week are the same. My hypothesis would be stated as:

$H_0: p_1 = p_2$

where:

p_1 = the proportion of 13- to 19-year-old males who use IMs every week

p_2 = the proportion of 13- to 19-year-old females who use IMs every week

The following table summarizes the data from the IM samples:

What can we conclude at the $\alpha = 0.10$ level?

Our sample proportion of male IM users, p_1 , and female users, p_2 , can be found by:

$\bar{x}_1 = 207, n_1 = 300, \bar{x}_2 = 266, n_2 = 350$

$p_1 = \frac{207}{300} = 0.69$ and $p_2 = \frac{266}{350} = 0.76$

$n_1 \geq 300, n_2 \geq 350$

To determine the calculated z-score, we need to know the standard error of the difference between two proportions (that's a mouthful), SE , which is found using:

$SE = \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

Our problem is that we don't know the values of p_1 and p_2 , the actual population proportions of male and female IM users. The next best thing is to calculate the estimated standard error of the difference between two proportions, SE , using the following equation:

$SE = \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

where p , the estimated overall proportion of two populations, is found using the following equation:

$$\hat{p} = \frac{207 + 266}{300 + 350} = 0.728$$

For our IM example, the estimated standard error of the difference between two proportions is:

$$(0.728)(1-0.728)(\frac{1}{300} + \frac{1}{350}) = 0.035$$

Bob's Basics

The term $(p_1 - p_2)$ refers to the hypothesized difference between the two population proportions. When the null hypothesis is testing that there is no difference between population proportions, then the term $(p_1 - p_2)$ is set to 0.



Now we can finally determine the calculated z-score using:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\text{SE}}$$

$$z =$$

For the IM example, our calculated z-score becomes: $\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\text{SE}} = \frac{(0.69 - 0.76) - 0}{0.035}$

$$= -2.00$$

$$= -2.00$$

The critical z-scores for a two-tail test with $\alpha = 0.10$ are +1.64 and -1.64. This hypothesis test is shown graphically in Figure 17.9.

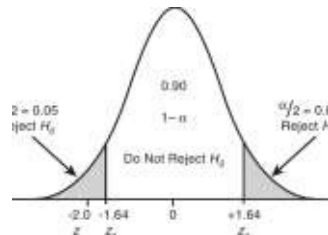
268 Part J: Inferential Statistics

As you can see in Figure 17.9, the calculated z-score of -2.00 is within the "Reject H_0 " region. There, we conclude that the proportions of male and female IM users between 13 and 19 years old are not equal to each other.

Figure 17.9

Hypothesis test for the IM example.

Hypothesis Test for the Difference in Proportions



$$\alpha = 0.05$$

Reject H_0

$$\alpha/2 = 0.05$$

Reject H_0

Number of Standard Deviations from the Mean



Bob's Basics

The standard error of the difference between two proportions describes the variation in the difference between two sample proportions and is calculated using

a -

$P_1 - P_2$

$-P_1, P_2 - P_2$

The estimated standard error of the difference

between two proportions approximates the variation in the difference between two

sample proportions and is calculated using $\hat{p} = \frac{\sum P_j}{n}$. The estimated overall proportion of two populations is the weighted average of two sample proportions and is calculated using $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$.

The p-value for these samples can be found using the normal z-score table found in Appendix B as follows:

$$2(P[z > +2.00]) = 2(1 - P[z < +2.00])$$

$$2(P[z > +2.00]) = 2(1 - 0.9772) = 0.0456$$

This completes our invigorating journey through the land of hypothesis testing. Don't be too sad, though. We'll have the pleasure of revisiting this technique in Part 4 of this book—Advanced Inferential Statistics. I just bet you can't wait.

Your Turn

1. Test the hypothesis that the average SAT math scores from students in Pennsylvania and Ohio are different. A sample of 45 students from Pennsylvania had an average score of 552, whereas a sample of 38 Ohio students had an average score of 530. Assume the population standard deviations for Pennsylvania and Ohio are 105 and 114 respectively. Test at the $\alpha = 0.05$ level. What is the p-value for these samples?

2. A company tracks satisfaction scores based on customer feedback from individual stores on a scale of 0 to 100. The following data represents the customer scores from Stores 1 and 2.

Store 1:

90 87 93 75 88 96 90 82 95 97 78

Store 2:

82 85 90 74 80 89 75 81 93 75

Assume population standard deviations are equal but unknown and that the population is normally distributed. Test the hypothesis using $\alpha =$

0.10.

3. A new diet program claims that participants will lose more than 15 pounds after completion of the program. The following data represents the before and after weights of nine individuals who completed the program. Test the claim at the

$\alpha = 0.05$ level.

Before: 221 215 206 185 202 197 244 188 218 After: 200 192 195 166 187
177 227 165 201

4. Test the hypothesis that the proportion of home ownership in the state of Florida exceeds the national proportion at the $\alpha = 0.01$ level using the following data.

270 Part 3: Inferential Statistics

Population Number of Successes Sample Size

Florida 272 400

Nation 390 600

What is the p-value for these samples? 5. Test the hypothesis that the average hourly wage for City A is more than \$0.50 per

Test at the $\alpha = 0.05$ level. What is the /i-value for this test?

6. Test the hypothesis that the average number of days that a home is on the market in City A is different from City B using the following sample data:

City A: 12 8 19 10 26 4 15 20 18 25 7 11

City B: 15 31 14 5 18 20 10 7 25 20 27

Assume population standard deviations are unequal and that the population is normally distributed. Test the hypothesis using $\alpha = 0.10$.

The Least You Need to Know

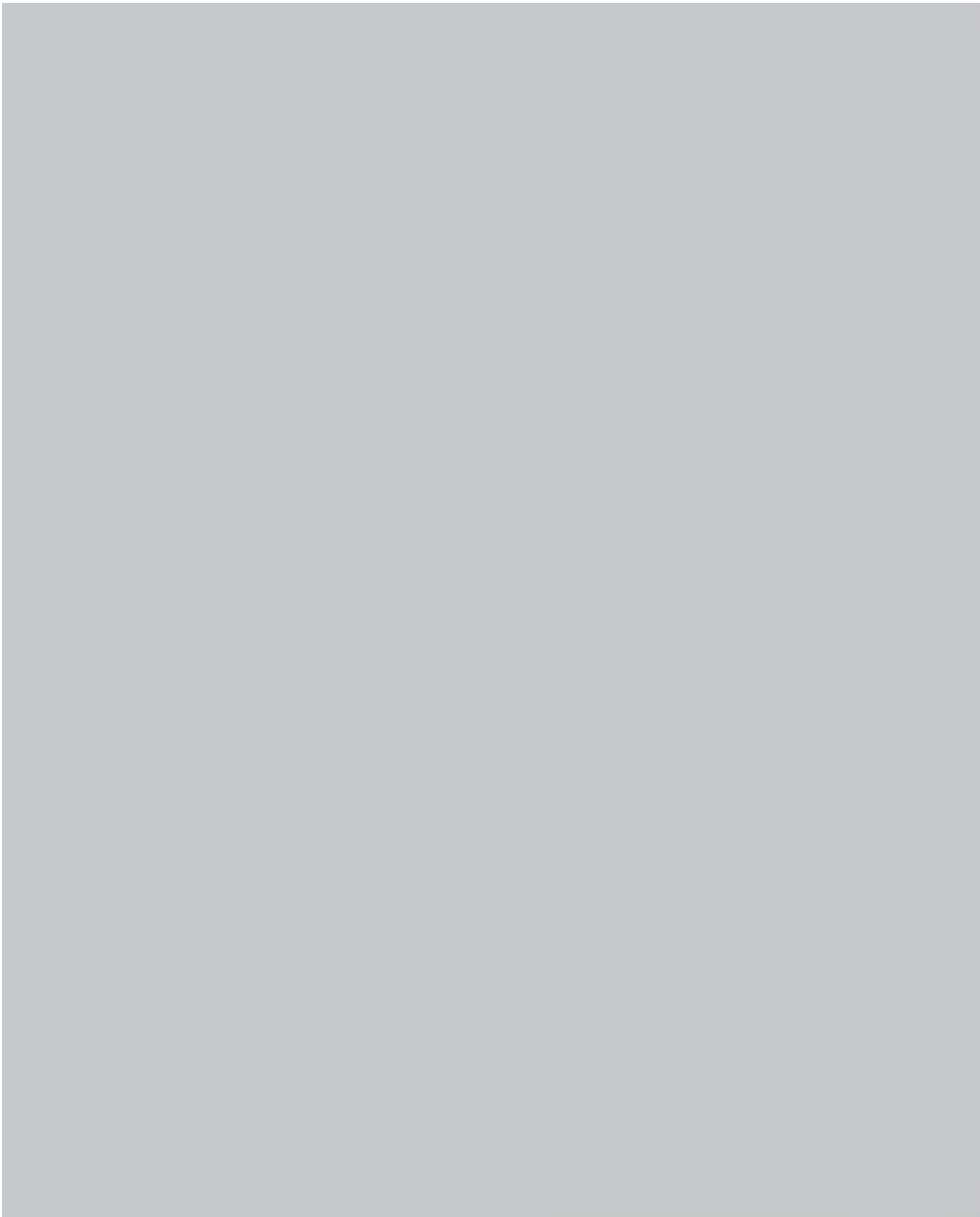
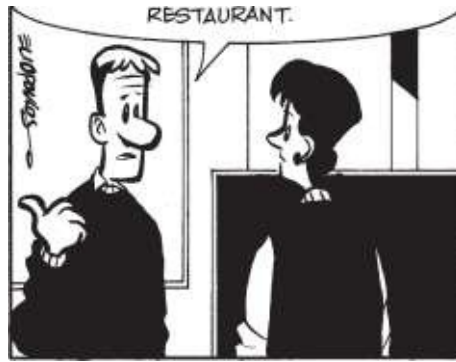
- ◆ We use the normal distribution for the hypothesis test for the difference between means when $n > 30$ for both samples.
- ◆ We use the normal distribution for the hypothesis test for the difference between means when $n < 30$ for either sample, if σ , and C , are known, and both populations are normally distributed.
- ◆ We use the Student's t -distribution for the hypothesis test for the difference between means when $n < 30$ for either sample, σ , and C , are unknown, and both populations are normally distributed.
- ◆ With dependent samples, the observation from one sample is related to an observation from another sample. With independent samples, there is no relationship in the observations between the samples.

Part 4

Advanced Inferential Statistics

We covered a lot of ground so far in the first three parts of this book. What could possibly be left? Well, the last few topics focus on the more advanced statistical methods (don't worry, you can handle it) of chi-square tests, analysis of variance, and simple regression. Armed with these techniques, we can determine whether two categorical variables are related (chi-square), compare three or more populations (analysis of variance), and describe the strength and direction of the relationship between two variables (simple regression). After you have mastered these concepts, the sky is the limit!

I WANTED TO DO THE CHI-SQUARE TEST, BUT I COULDN'T SEE IT ON THE MENU AT MY LOCAL CHINESE



(hapte

18

The Chi-Square Probability Distribution

In This Chapter

- ◆ Performing a goodness-of-fit test with the chi-square distribution
- ◆ Performing a test of independence with the chi-square distribution
- ◆ Using contingency tables to display frequency distributions

In the last three chapters, we explored the wonderful world of hypothesis testing as we compared means and proportions of one and two populations, making an educated conclusion about our initial claims. With that technique under our belt, we are now ready for bigger and better things.

In this chapter, we will compare two or more proportions using a new probability distribution: the chi-square. With this new test, we can confirm whether a set of data follows a specific probability distribution, such as the binomial or Poisson. (Remember those? They're back!) We can also use this distribution to determine whether two variables are statistically independent. It's actually a lot of fun—really it is!

Review of Data Measurement Scales

In Chapter 2, we discussed the different type of data measurement scales, which were nominal, ordinal, interval, and ratio. Here is a brief refresher of each:

- ◆ Nominal level of measurement deals strictly with qualitative data. Observations are simply assigned to predetermined categories. One example is gender of the respondent with the categories being male and female.

◆ Ordinal measurement is the next level up. It has all the properties of nominal data with the added feature that we can rank order the values from highest to lowest. An example would be ranking a movie as great, good, fair, or poor.

◆ Interval level of measurement involves strictly quantitative data. Here we can use the mathematical operations of addition and subtraction when comparing values. For this data, the difference between the different categories can be measured with actual numbers and also provides meaningful information. Temperature measurement in degrees Fahrenheit is a common example here.

◆ Ratio level is the highest measurement scale. Now we can perform all four mathematical operations to compare values. Examples of this type of data are age, weight, height, and salary. Ratio data has all the features of interval data with the added benefit of a "true zero point," meaning that a zero data value indicates the absence of the object being measured.

def'i'oi'tiein

The chi-square distribution

is used to perform hypothesis testing on nominal and ordinal data.

The hypothesis testing that we covered in the last three chapters strictly used interval and ratio data. However, the chi-square distribution in this chapter will allow us to perform hypothesis testing on nominal and ordinal data.

The two major techniques that we will learn about are using the chi-square distribution to perform a goodness-of-fit test and to test for the independence of two variables. So let's get started!

The Chi-Square Goodness-of-Fit Test

One of the many uses of the chi-square distribution is to perform a goodness-of-fit test, which uses a sample to test whether a frequency distribution fits the predicted

distribution. As an example, let's say that a new movie in the making has

an expected distribution of ratings summarized in the following table.

Expected Movie-Rating Distribution

Number of Stars

Percentage

5 4 3 2 1 Total

40% 30%

20% 5% 5%

100%

After its debut, a sample of 400 moviegoers were asked to rate the movie, with the results shown in the following table.

Observed Movie-Rating Distribution

Number of Stars

Number of Observations

5

4 3

2 1 Total

145 128

73

n

22 400

Can we conclude that the expected movie ratings are true based on the observed ratings of 400 people?

def'i'ni'tiem

The goodness-of-fit test uses a sample to test whether a frequency distribution fits the predicted distribution.

276 Parfi: Advanced Inferential Statistics

Stating the Null and Alternative Hypothesis

The null hypothesis in a chi-square goodness-of-fit test states that the sample of observed frequencies supports the claim about the expected frequencies. The alternative hypothesis states that there is no support for the claim pertaining to the expected

frequencies. For our movie example, the hypothesis statement would look like the following:

Bob's Basics



The total number of expected [E] frequencies must be equal to the total number of observed (O) frequencies.

H_0 : The actual rating distribution can be described by the expected distribution.

H_a : The actual rating distribution differs from the expected distribution.

We will test this hypothesis at the $\alpha = 0.10$ level.

Observed Versus Expected Frequencies

The chi-square test basically compares the observed (O) and expected (E) frequencies to determine whether there is a statistically significant difference. For our movie example, the observed frequencies are simply the number of observations collected for each category of our sample. The expected frequencies are the expected number of observations for each category and are calculated in the following table.

defi'iortieio

Observed frequencies are the number of actual observations noted for each category of a frequency distribution with chi-squared analysis.

Expected frequencies are the number of observations that would be expected for each category of a frequency distribution assuming the null hypothesis is true with chi-squared analysis.

Expected Frequency Table

Movie Expected Sample Expected Observed

Rating Percentage Size Frequency (E) Frequency (O)

We are now ready to calculate the chi-square statistic.

Calculating the Chi-Square Statistic

The chi-square statistic is found using the following equation:

where:

O = the number of observed frequencies for each category E = the number of expected frequencies for each category

The calculation using this equation is shown in the following table.

The Calculated Chi-Square Score for the Movie Example

Determining the Critical Chi-Square Score

The critical chi-square score, X_{α} , depends on the number of degrees of freedom, which for this test would be:

$d.f. = k - 1$

278 Part 4: Advanced Inferential Statistics

where k equals the number of categories in the frequency distribution. For the movie example, there are 5 categories, so $d.f. = 5 - 1 = 4$.

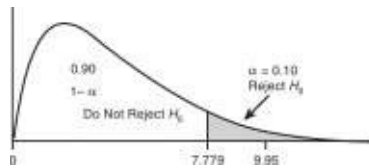
The critical chi-square score is read from the chi-square table found on Table 5 in Appendix B of this book. Here is an excerpt of this table.

Critical Chi-Square Values

For $\alpha = 0.10$ and d.f. = 4, the critical chi-square score, X_{α} = 7.779, is indicated in the underlined part of the table. Figure 18.1 shows the results of our hypothesis test.

Figure 18.1

Chi-square test for the movie example.



$\alpha = 0.10$ Reject H_0

According to Figure 18.1, the calculated chi-square score of 9.95 is within the "Reject H_0 " region, which leads us to the conclusion that the actual movie-rating frequency distribution differs from the expected distribution. We will always reject H_0 as long as $X > x_{\alpha}$

Also, because the calculated chi-square score for the goodness-of-fit test can only be positive, the hypothesis test will always be a one-tail with the rejection region on the right side.

Using Excel's CHIINV Function

You don't have a chi-square distribution table handy? No need to panic. We can generate critical chi-square scores using Excel's CHIINV function, which has the following characteristics:

CHIINV(probability, deg-freedom)

where:

probability = the level of significance, a deg-freedom = the number of degrees of freedom

For instance, Figure 18.2 shows the CHIINV function being used to determine the critical chi-square score for $\alpha = 0.10$ and d.f. = 4 from our previous example.

Figure 18.2

Excel's CHIINV function.

Cell A1 contains the Excel formula =CHIINV(0.10, 4) with the result being 7.779. This probability is underlined in the previous table.

Characteristics of a Chi-Square Distribution

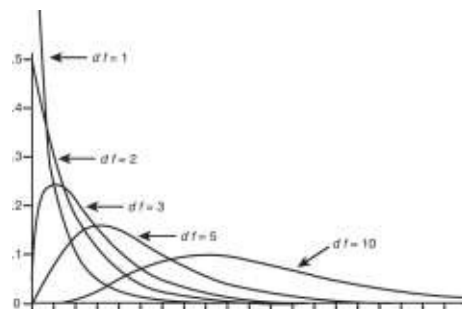
We can see from Figure 18.2 that the chi-square distribution is not symmetrical but rather has a positive skew. The shape of the distribution will change with the number of degrees of freedom as shown in Figure 18.3.

As the number of degrees of freedom increases, the shape of the chi-square distribution becomes more symmetrical.

Z80 Part 4: Advanced Inferential Statistics

Figure 18.3

Family of chi-square distributions.



i — T i i i i I i I i i i i I i i r

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Chi-Square Values

A Goodness-of-Fit Test with the Binomial Distribution

In past chapters, we have occasionally made assumptions that a population follows a specific distribution such as the normal or binomial. In this section, we can demonstrate how to verify this claim.

As an example, suppose that a certain major league baseball player claims the probability that he will get a hit at any given time is 30 percent. The following table is a frequency distribution of the number of hits per game over the last 100 games. Assume he has come to bat four times in each of the games.

Data for the Baseball Player

Number of Hits

Number of Games

0 1 2 3 4 Total

26

34

30

7

3

In other words, in 26 games he had 0 hits, in 34 games he had 1 hit, etc. Test the claim that this distribution follows a binomial distribution with $p = 0.30$ using $\alpha = 0.05$.

The hypothesis statement would look like the following:

H_0 : The distribution of hits by the baseball player can be described with the binomial probability distribution using $p = 0.30$.

H_1 : The distribution differs from the binomial probability distribution using $\pi = 0.30$.

Our first step is to calculate the frequency distribution for the expected number of hits per game. To do this, we need to look up the binomial probabilities in Table 1 from Appendix B for $n = 4$ (the number of trials per game) and $\pi = 0.30$ (the probability of a success). These probabilities, along with the calculations for the expected frequencies, are shown in the following table.

Expected Frequency Calculations for Baseball Player

Number of Hits per Game

Binomial Probabilities

Number of Games

Expected Frequency

0 1 2 3 4 Total

Before continuing, we need to make one adjustment to the expected frequencies. When using the chi-square test, we need at least five observations in each of the expected frequency categories. If there are less than five, we need to combine categories. In the previous table, we will combine 3 and 4 hits per game into one category to meet this requirement.



Bob's Basics

Expected frequencies do not have to be integer numbers because they only represent theoretical values.

282 Part 4: Advanced Inferential Statistics

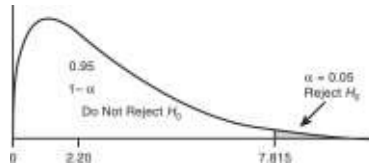
Now we are ready to determine the calculated chi-square score using the following table:

The Calculated Chi-Square Score for the Baseball Example

According to Table 5 in Appendix B, the critical chi-square score for $\alpha = 0.05$ and $d.f. = \mathcal{E} = 1 = 4 - 1 = 3$ is 7.815. This test is shown in Figure 18.4.

Figure 18.4

Chi-square test for the baseball example.



According to Figure 18.4, the calculated chi-square score of 2.20 is within the "Do Not Reject H_0 " region, which leads us to the conclusion that the baseball player's hitting distribution can be described with the binomial distribution using $\hat{p} = 0.30$.

Chi-Square Test for Independence

In addition to the goodness-of-fit test, the chi-square distribution can also test for independence between variables. To demonstrate this technique, I'm going to revisit the tennis example from Chapter 7.

If you recall, Debbie felt that a short warm-up period before playing our match was hurting her chances of beating me. After examining the conditional probabilities, I had to admit there was some evidence supporting Debbie's claim. However, I'm not one to take this sitting down. I demand justice, I demand further evidence, I demand a recount. (Oh, wait a minute, this isn't Florida.) I demand ... a hypothesis test using the chi-square distribution!

Unbeknownst to Debbie, I have meticulously collected data from our 50 previous matches. The following table represents the number of wins for each of us according to the length of the warm-up period.

Observed Frequencies for Tennis Example

This is known as a contingency table, which shows the observed frequencies of two variables. In this case, the variables are warm-up time

and tennis player. The table is organized into r rows and c columns. For our table, $r = 2$ and $c = 3$. An intersection of a row and column is known as a cell. A contingency table has $r \cdot c$ cells, which in our case, would be 6.

The chi-square test of independence will

determine whether the proportion of times that Debbie wins is the same for all three warm-up periods. If the outcome of the hypothesis test is that the proportions are not the same, we conclude that the length of warm-up does impact the performance of the players. But I have my doubts.

First we state the hypotheses as:

H_0 : Warm-up time is independent of performance

H_a : Warm-up time affects performance

A contingency table shows the observed frequencies of two variables. An intersection of a row and column in a contingency table is known as a cell. A contingency table has $r \cdot c$ cells.

We will test this hypothesis at $\alpha = 0.10$ level.

Our next step is to determine the expected frequency of each cell in the contingency table under the assumption that the two variables are independent. We do this using the following equation:

„ Total of Row r ■ Total of Column c Total Number of Observations

where E_{rt} - the expected frequency of the cell that corresponds to the intersection of Row r and Column c .

The following table applies this notation to our tennis example.

Total Observations

The total number of observations for this example is 50, which we can confirm by adding $23+27$ or $18 +19 +13$. We can now determine the expected frequencies for each cell:

$$E_{11} = \hat{M} = 8.28 \text{ £}, E_{12} = \hat{M} = 8.74 \text{ £}, E_{13} = \hat{M} = 5.98$$

$$E_{21} = \hat{M} = 9.72 \text{ £}, E_{22} = \hat{M} = 10.26 \text{ £}, E_{23} = \hat{M} = 7.02$$

$$E_{31} = \hat{M} = 9.72 \text{ £}, E_{32} = \hat{M} = 10.26 \text{ £}, E_{33} = \hat{M} = 7.02$$

$$E_{41} = \hat{M} = 9.72 \text{ £}, E_{42} = \hat{M} = 10.26 \text{ £}, E_{43} = \hat{M} = 7.02$$

The following table summarizes these findings.

Expected Frequencies for Tennis Example

To determine the critical chi-square score, we need to know the number of degrees of freedom, which for the independence test would be:

$$d.f. = (r-1)(c-1)$$

For this example, we have $(r-1)(c-1) = (2-1)(3-1) = 2$ degrees of freedom.

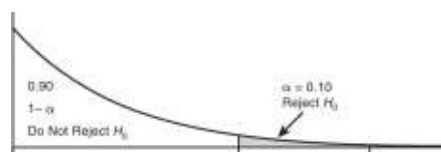
According to Table 5 in Appendix B, the critical chi-square score for $\alpha = 0.10$ and $d.f. = 2$ is 4.605. This test is shown in Figure 18.5.

According to Figure 18.5, the calculated chi-square score of 7.25 is within the "Reject H_0 " region, which leads us to the conclusion that there is a relationship between warm-up time and performance when Debbie and I play tennis. Darn it—once again, Debbie is right. Boy, does that have a familiar ring to it.

286 Part k: Advanced Inferential Statistics

Figure 18.5

Chi-square test for the tennis example.



$\alpha = 0.10$ Reject H_0

However, I do have one consolation. The chi-square test of independence only investigates whether a relationship exists between two variables. It does not conclude anything about the direction of the relationship. In other words, from a statistical perspective, Debbie cannot claim that she is disadvantaged by the short warm-up time. She can only claim that warm-up time has some effect on her performance. We statisticians always leave ourselves a way out!

Your Turn

1. A company believes that the distribution of customer arrivals during the week are as follows:

Day

Expected Percentage of Customers

Monday

Tuesday

Wednesday

Thursday

Friday

Saturday

Total

10 10 15 15 20 30 100

A week was randomly chosen and the number of customers each day was counted. The results were: Monday—31, Tuesday—18, Wednesday—36, Thursday—23, Friday—47, Saturday—60. Use this sample to test the expected distribution using $\alpha = 0.05$.

2. An e-commerce site would like to test the hypothesis that the number of hits per minute on their site follows the Poisson distribution with $X =$

3. The following data was collected:

Number of Hits

Per Minute 0 1 2 3 4 5 6 7 or More

Frequency 22 51 72 92 60 44 25 14

Test the hypothesis using $\alpha = 0.01$.

3. An English professor would like to test the relationship between an English grade and the number of hours per week a student reads. A survey of 500 students resulted in the following frequency distribution.

Test the hypothesis using $\alpha = 0.05$.

John Armstrong, salesman for the Dillard Paper Company, has five accounts to visit each day. It is suggested that the random variable, successful sales visits by Mr. Armstrong, may be described by the binomial distribution, with the probability of a successful visit being 0.4. Given the following frequency distribution of Mr. Armstrong's number of successful sales visits per day, can we conclude that the data actually follows the binomial distribution? Use $\alpha = 0.05$.

288 Part k: Advanced Inferential Statistics

The Least You Need to Know

- ◆ The chi-square distribution is not symmetrical but rather has a positive skew. As the number of degrees of freedom increases, the shape of the chi-square distribution becomes more symmetrical.
- ◆ The chi-square distribution allows us to perform hypothesis testing on nominal and ordinal data.
- ◆ We can use the chi-square distribution to perform a goodness-of-fit test, which uses a sample to test whether a frequency distribution fits a predicted distribution.
- ◆ The chi-square test for independence investigates whether a

relationship exists between two variables. It does not, however, test the direction of that relationship.

A contingency table shows the observed frequencies of two variables. An intersection of a row and column in a contingency table is known as a cell.

(hapte

Analysis of Variance



In This Chapter

- ◆ Comparing three or more population means using analysis of variance (ANOVA)
- ◆ Using the F-distribution to perform a hypothesis test for ANOVA
- ◆ Using Excel to perform a one-way ANOVA test
- ◆ Comparing pairs of sample means using the Scheffe test

In Chapter 17, you learned about a hypothesis test where you could compare the means of two different populations to see whether they were different. But what if you want to compare the means of three or more populations? Well, you've come to the right place, because that's what this chapter is all about.

To perform this new hypothesis test, I need to introduce one last probability distribution, known as the F-distribution. The test that we will perform has a very impressive name associated with it—the analysis of variance. This test is so special, it even has its own acronym: ANOVA. Sounds like something from outer space ... keep reading to find out.

290 Part k: Advanced Inferential Statistics

One-Way Analysis of Variance

If you want to compare the means for three or more populations, ANOVA is the test for you. Let's say I'm interested in determining whether there is a difference in consumer satisfaction ratings between three fast-food chains. I would collect a sample of satisfaction ratings from each chain and test to see whether there is a significant difference between the sample means. Suppose my data look like the following:

Population	Fast-Food Chain	Sample Mean Rating
1	McDoogles	7.8
2	Burger Queen	8.2
3	Windy's	8.3

My hypothesis statement would look like the following: $H_0: \mu_1 = \mu_2 = \mu_3$
 H_a : not all μ 's are equal

Essentially, I'm testing to see whether the variations in customer ratings from the previous table are due to the fast-food chains or whether the variations are purely random. In other words, do customers perceive any differences in satisfaction between the three chains? If I reject the null hypothesis, however, my only conclusion is that a difference does exist. Analysis of variance does not allow me to compare population means to one another to determine which is greater. That task requires further analysis.

Bob's Basics



To use one-way ANOVA, the following conditions must be present:

- ♦ The populations of interest must be normally distributed.
- ♦ The samples must be independent of each other.

◆ Each population must have the same variance.

A factor in ANOVA describes the cause of the variation in the data. In the previous example, the factor would be the fast-food chain. This would be considered a one-way ANOVA because we are considering only one factor. More complex types of ANOVA can examine multiple factors, but that topic goes beyond the scope of this book.

A level in ANOVA describes the number of categories within the factor of interest. For our example, we have three levels based on the three different fast-food chains being examined.

To demonstrate one-way ANOVA, I'll use the following example. I admit, much to Debbie's chagrin, that I am clueless when it comes to lawn care. My motto is, "If it's green, it's good." Debbie, on the other hand, knows exactly what type of fertilizer to get and when to apply it during the year. I hate spreading this stuff on the lawn because it apparently makes the grass grow faster, which means I have to cut it more often.

def

•rnrticn

To make matters worse, we have a neighbor, Bill, whose yard puts my yard to shame. Mr. "Perfect Lawn" is out every weekend, meticulously manicuring his domain until it looks like the home field for the National Lawn Bowling Association. This gives Debbie a serious case of "lawn envy." Bill even has one of those cute little carts that he pulls on the back of his tractor. I asked Debbie if I could get one for my tractor, but she said based on my "Lawn IQ" I would probably just injure myself.

Anyway, there are several different types of analysis of variance, and covering them all would take a book unto itself. So throughout the remainder of this chapter, we'll use my lawn-care topic to describe two basic ANOVA procedures.

A factor in ANOVA describes the cause of the variation in the data. When only one factor is being considered, the procedure is known as one-way ANOVA. A level in ANOVA describes the number of categories within the factor of interest.

Completely Randomized ANOVA

The simplest type of ANOVA is known as completely randomized one-way ANOVA, which involves an independent random selection of observations for each level of one factor. Now that's a mouthful! To help explain this, let's say I'm interested in comparing the effectiveness of three lawn fertilizers. Suppose I select 18 random patches of my precious lawn and apply either Fertilizer 1, 2, or 3 to each of them. After a week, I mow the patches and weigh the grass clippings.

def'i'ni'tien

The simplest type of ANOVA is known as completely randomized one-way ANOVA, which involves an independent random selection of observations for each level of one factor.

292 Part 4: Advanced Inferential Statistics

The factor in this example is fertilizer. There are three levels, representing the three types of fertilizer we are testing. The table that follows indicates the weight of the clippings in pounds from each patch. The mean and variance of each level are also shown.

Data for Lawn Clippings

We'll refer to the data for each type of fertilizer as a sample. From the previous table, we have three samples, each consisting of six observations. The hypotheses statement can be stated as:

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_1: \text{not all } \mu_i \text{'s are equal}$

where μ_1 , μ_2 , and μ_3 are the true population means for the pounds of grass clippings for each type of fertilizer.

Partitioning the Sum of Squares

The hypothesis test for ANOVA compares two types of variations from the samples. We first need to recognize that the total variation in the data

from our samples can be divided, or as statisticians like to say, "partitioned," into two parts.

The first part is the variation within each sample, which is officially known as the sum of squares within (SSW). This can be found using the following equation:

k

where k = the number of samples (or levels). For the fertilizer example, $k = 3$ and: $\sum (x_{ij} - \bar{x}_j)^2 = 1.01 + 1.70 + 0.96$

$\sum (x_{ij} - \bar{x}_j)^2 = 3.67$

The sum of squares within can now be calculated as:

$$SSW = (3 - 1)1.01 + (3 - 1)1.70 + (3 - 1)0.96 = 8.35$$

Some textbooks will also refer to this value as the error sum of squares (SSE).

The second partition is the variation among the samples, which is known as the sum of squares between (SSB). This can be found by:

SSB

$\sum n_j(\bar{x}_j - \bar{x})^2$

where \bar{x} is the grand mean or the average value of all the observations. For the fertilizer example:

Random Thoughts

Some textbooks will also refer to this SSB value as the treatment sum of squares (SSTR).

$\sum x_i^2 = 9.12$

$\sum x_i = 10.92$

$\bar{x} = 3.64$

We find \bar{x} , the grand mean, using:

$$\bar{x} = \frac{\sum x_i}{N}$$

$$\bar{x} = \frac{10.2 + 8.5 + 8.4 + 10.5 + \dots + 9.1 + 10.5 + 9.5}{18}$$

$$\bar{x} = 9.83$$

where N - the total number of observations from all samples. For the fertilizer example:

$$\bar{x} = \frac{10.2 + 8.5 + 8.4 + 10.5 + \dots + 9.1 + 10.5 + 9.5}{18}$$

$$\bar{x} = 9.83$$

$$\bar{x} = 9.83$$

We can now calculate the sum of squares between: $SSB = \sum_{j=1}^J n_j (\bar{x}_j - \bar{x})^2$

$$SSB = 6(9.12 - 9.83)^2 + 6(10.92 - 9.83)^2 + 6(9.48 - 9.83)^2 = 10.86$$

Random Thoughts

ANOVA does not require that all the sample sizes are equal, as they are in the fertilizer example. See Problem 1 in the "Your Turn" section as an example of unequal sample sizes.

Finally, the total variation of all the observations is known as the total sum of squares (SST) and can be found by:

$$SST = \sum_{i=1}^N (x_i - \bar{x})^2$$

This equation may look nasty, but it is just the difference between each observation and the grand mean squared and then totaled over all of the observations. This is clarified more in the following table.

$$SST = \sum_{i=1}^N (x_i - \bar{x})^2$$

$$SST = \sum_{i=1}^N (x_i - \bar{x})^2$$

0.14 1.77 2.04 0.45 0.69 2.99 3.13 4.71 0.40 0.22 0.01 7.13 2.99 0.69 0.76
0.53 0.45 0.11

$$SST = \sum (x_i - \bar{x})^2 = 29.21$$

'??(*■

This total sum of squares calculation can be confirmed recognizing that:
 $SST = SSIV + SSB$ $SST = 18.35 + 10.86 = 29.21$

Note that we can determine the variance of the original 18 observations, s^2 , by:

$$s^2 = \frac{SST}{n-1} = \frac{29.21}{18-1}$$

$$s^2 = 1.72$$

$$s = 1.31$$

This result can be confirmed by using the variance equation that we discussed in

Chapter 5 or by using Excel.

Determining the Calculated F-Statistic

To test the hypothesis for ANOVA, we need to compare the calculated test statistic to a critical test statistic using the F-distribution. The calculated F-statistic can be found using the equation: $F = \frac{MSB}{MSW}$ where MSB is the mean square between, found by:

$$MSB = \frac{SSB}{k-1}$$

and MSW is the mean square within, found by: $MSW = \frac{SSW}{N-k}$

N-k Now, let's put these guys to work with our fertilizer example.

$$MSB = \frac{10.86}{3-1} = 5.43$$

$$MSW = \frac{18.35}{18-3} = 1.22$$

„MSB 5.43 „ „ r ■ 4.45

MSW 1.22

If the variation between the samples (MSB) is much greater than the variation within the samples (MSW), we will tend to reject the null hypothesis and conclude that there is a difference between population means. To complete our test for this hypothesis, we need to introduce the F-distribution.

296 Part k: Advanced Inferential Statistics

Bob's Basics

» \ " ^ e mean s q uare between (MSB) is a measure of variation between the sample // \ \ means - The mean square within (MSW) is a measure of variation within each

sample. A large MSB variation, relative to the MSW variation, indicates that the sample means are not very close to one another. This condition will result in a large value of F, the calculated F-statistic. The larger the value of F, the more likely it will exceed the critical F-statistic (to be determined shortly), leading us to conclude there is a difference between population means.



Determining the Critical F-Statistic

We use the F-distribution to determine the critical F-statistic, which is compared to the calculated F-statistic for the ANOVA hypothesis test. The critical F-statistic, F_{α, v_1, v_2} , depends on two different degrees of freedom, which are determined by:

$v_1 = k - 1$ and $v_2 = N - k$

For our fertilizer example:

$v_1 = 3 - 1 = 2$ and $v_2 = 18 - 3 = 15$

The critical F-statistic is read from the F-distribution table found in Table 6 in Appendix B of this book. Here is an excerpt of this table.

Table of Critical F-Statistics

$\alpha = 0.05$

$\backslash v_1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10$

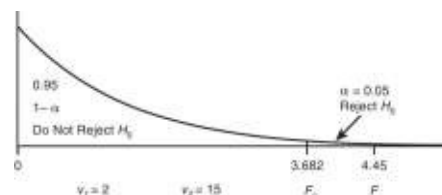
1 161.448 199.500 215.707 224.583 230.162 233.986 236.768 238.882
240.543 241.882

Note that this table is based only on $\alpha = 0.05$. Other values of α will require a different table. For $v_1 = 2$ and $v_2 = 15$, the critical F-statistic, $F_{0.05, 2, 15} = 3.682$, as indicated in the underlined part of the table. Figure 19.1 shows the results of our hypothesis test.

Figure 19.1

ANOVA test for the fertilizer

example, $\alpha = 0.05$ Reject H_0



According to Figure 19.1, the calculated F-statistic of 4.45 is within the "Reject H_0 " region, which leads us to the conclusion that the population means are not equal. We will always reject H_0 as long as $F_{obs} > F_c$.



n

Bob's Basics

The F-distribution has the following characteristics:

It is not symmetrical but rather has a positive skew.

◆ The shape of the F-distribution will change with the degrees of freedom specified by the values of v_1 and v_2 .

◆ As v_1 and v_2 increase in size, the shape of the F-distribution becomes more symmetrical.

■ The total area under the curve is equal to 1.

◆ The F-distribution mean is approximately equal to 1.

298 Part k: Advanced Inferential Statistics

Our final conclusion is that one or more of those darn fertilizers is making the grass grow faster than the others. Sounds like trouble to me.

It

, Wrong Number

/+> '^p Even though we have rejected H_0 , and concluded that the population means are not all equal, ANOVA does not allow us to make comparisons between means. In other words, we do not have enough evidence to conclude that Fertilizer 2 produces more grass clippings than Fertilizer 1. This requires another test known as pair-wise comparisons, which we'll address later in this chapter.

Now let's explore how Excel can take some of the burden from all these nasty calculations.

Using Excel to Perform One-Way ANOVA

I'm sure you've come to the conclusion that calculating ANOVA manually is a lot of work, and I think you'll be amazed how easy this procedure is using Excel.

1. Start by placing the fertilizer data in Columns A, B, and C in a blank sheet.

2. Go to the Tools menu and select Data Analysis. (Refer to the section "Installing the Data Analysis Add-in" from Chapter 2 if you don't see the Data Analysis command on the Tools menu.)

3. From the Data Analysis dialog box, select Anova: Single Factor as shown in Figure 19.2 and click OK.

Figure 19.2

Setting up the one-way ANOVA in Excel.

ANOVA

ANOVA: Single Factor

Input Range: \$A\$1:\$A\$10

By Inputing the above information, the ANOVA results are displayed as follows.

ANOVA for Factor A

4. Set up the Anova: Single Factor dialog box according to Figure 19.3.

5. Click OK. Figure 19.4 shows the final ANOVA results.

Figure 19.4

Final results of the one-way ANOVA in Excel.

These results are consistent with what we found doing it the hard way in the previous sections. Notice that the Rvalue = 0.0305 for this test, meaning we can reject H_0 because this p-value < α . If you remember, we had set $\alpha = 0.05$ when we stated the hypothesis test.

Pairwise Comparisons

Once we have rejected H_0 using ANOVA, we can determine which of the sample means are different using the Scheffe test. This test compares each pair of sample means from the ANOVA procedure. For the fertilizer example, we would compare X_1 versus X_2 , X_1 versus X_3 , and X_2 versus X_3 to see whether any differences exist.

300 Part 4: Advanced Inferential Statistics

Bob's Basics



After rejecting H_0 using χ^2 ANOVA, we can determine

which of the sample means are different using the Scheffe test.

First, the following test statistic for the Scheffe test, F_s , is calculated for each of the pairs of sample means:

where: \bar{x}_i , X_i - the sample means being compared SSW = the sum of squares within from the ANOVA procedure n_i , n_b - the samples sizes k - the number of samples (or levels)

Comparing \bar{x}_t and \bar{x}_2 , we have: $(9.12 - 10.92)^2$

$F_c =$

18.35

$5 + 5 + 5$

11

$-\frac{1}{6} + -\frac{1}{6}$

$3.24 - 1.22[0.33]$

8.048

Comparing \bar{x}_t and X_i , we have:

$(9.12 - 9.48)^2 - 0.13$

18.35

11

- + -6 6

1.22[0.33]

: 0.323

5 + 5 + 5 Comparing \bar{x}_2 and \bar{x}_i , we have:

$\sqrt{(10.92-9.48)^2} = 2.07$

18.35

5 + 5 + 5

1 1

- + -6 6

1.22[0.33]

= 5.142

Next the critical value for the Scheffe test, F_{sc} , is determined by multiplying the critical F-statistic from the ANOVA test by $k - 1$ as follows:

For the fertilizer example: $F_{ma,u} = 3.682$ $F_{sc} = (3 - 1)(3.682) = 7.364$

If $F_s < F_{sc}$, we conclude there is no difference between the sample means; otherwise there is a difference. The following table summarizes these results.

Summary of the Scheffe Test

Sample Pair F. F_{sc} Conclusion

According to our results, the only statistically significant difference is between Fertilizer 1 and Fertilizer 2. It appears that Fertilizer 2 is more effective in making grass grow faster when compared to Fertilizer 1. I better keep Debbie away from this brand.

Completely Randomized Block ANOVA

Now let's modify the original fertilizer example: rather than select 18 random samples from my lawn, we are going to collect 3 random samples from 6 different lawns. Using the original data, the samples look as follows:

302 Part 4: Advanced Inferential Statistics

One concern in this scenario is that the variations in the lawns will account for some of the variation in the three fertilizers, which may interfere with our hypothesis test. We can control for this possibility by using a completely randomized block ANOVA, which is used in the previous table. The type of fertilizer is still the factor, and the lawns are called blocks.

definition

Completely randomized block ANOVA controls for variations from other sources than the factors of interest. This is accomplished by grouping the samples using a blocking variable.

There are two hypotheses for the completely randomized block ANOVA. The first (primary) hypothesis tests the equality of the population means, just like we did earlier with one-way ANOVA:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_j$$

H_a : not all μ 's are equal

The secondary hypothesis tests the effectiveness of the blocking variable as follows:

H_0 : the block means are all equal H_a : the block means are not all equal

The blocking variable would be an effective contributor to our ANOVA model if we can reject H_a and claim that the block means are not equal to each other.

Partitioning the Sum of Squares

For the completely randomized block ANOVA, the sum of squares total is partitioned into three parts according to the following equation:

$$SST = SSW + SSB + SSBL$$

where:

SSW - sum of squares within

SSB - sum of squares between

SSBL = sum of squares for the blocking variable (lawns)

Fortunately for us, the calculations for SST and SSW are identical to the one-way ANOVA procedure that we've already discussed, so those values remain unchanged (SST = 29.21 and SSB = 10.86). We can find the sum of squares block (SSBL) by using the equation:

SSBL =

$\sum_{j=1}^b (\bar{X}_j - \bar{X})^2$ where:

\bar{X}_j = the average observation of each blocking level

b - the number of blocking levels (b = 6 for our example)

Using the values from the previous table, we have:

$$SSBL = 3(9.97 - 9.83)^2 + 3(9.83 - 9.83)^2 + 3(9.43 - 9.83)^2 + 3(9.97 - 9.83)^2 + 3(9.80 - 9.83)^2 + 3(10.03 - 9.83)^2$$

SSBL = 0.72 That leaves us with the sum of squares within (SSW), which we can find using: $SSW = SST - SSB - SSBL$

$$SSW = 29.21 - 10.86 - 0.72 = 17.63$$

Almost done!

Determining the Calculated F-Statistic

Since we have two hypothesis tests for the completely randomized block

ANOVA, we have two calculated F-statistics. The F-statistic to test the equality of the population means (the original hypothesis) is found using:

MSB ~ MSW where MSB is the means square between, found by:

$$MSB = \frac{SSB}{k-1}$$

and MSW is the mean square within, found by:

$$MSW = \frac{SSW}{n-k}$$

$$MSW = \frac{SSW}{n-k}$$

IQk Part 4: Advanced Inferential Statistics

Inserting our fertilizer values into these equations looks like this:

$$MSB = \frac{SSB}{k-1} = \frac{10.5}{3-1}$$

$$MSB = \frac{10.5}{2} = 5.25$$

$$MSW = \frac{SSW}{n-k} = \frac{17.63}{12-3}$$

$$MSW = \frac{17.63}{9} = 1.96$$

$$F = \frac{MSB}{MSW} = \frac{5.25}{1.96} = 2.68$$

The second F-statistic will test the significance of the blocking variable (the second hypothesis) and will be denoted F'. We will determine this statistic using:

MSBL ~ MSW where MSBL is the (can you guess?) mean square blocking, found by:

$$MSBL = \frac{SSBL}{b-1}$$

$$MSBL = \frac{SSBL}{b-1}$$

$$MSBL = \frac{SSBL}{b-1}$$

MSW 1.76 We now need to sit back, catch our breath, and figure out what

all these numbers mean.

To Block or Not to Block, That Is the Question

First, we will examine the primary hypothesis, H_0 that all population means are equal using $\alpha = 0.05$. The degrees of freedom for this critical F-statistic would be:

$$\text{df} = k - 1 = 3 - 1 = 2$$

$$\text{df} = (k - 1)(b - 1) = (3 - 1)(6 - 1) = 10$$

The critical F-statistic from Appendix B is $F_{0.05, 2, 10} = 4.103$. Since the calculated F-statistic equals 3.09 and is less than this critical F-statistic, we fail to reject H_0 and cannot conclude that the fertilizer means are different.

We next examine the secondary hypothesis, H_0' , concerning the effectiveness of the blocking variable, also using $\alpha = 0.05$. The degrees of freedom for this critical F-statistic would be:

$$\text{df}' = k - 1 = 3 - 1 = 2$$

$$\text{df}' = (k - 1)(b - 1) = (3 - 1)(6 - 1) = 10$$

The critical F-statistic from Appendix B is $F_{0.05, 2, 10} = 3.326$. Since the calculated F-statistic, F , equals 0.08 and is less than this critical F-statistic, we fail to reject H_0' and cannot conclude that the block means are different.

What does all this mean? Since we failed to reject H_0' , the hypothesis that states the blocking means are equal, the blocking variable (lawns) proved not to be effective and should not be included in the model. Including an ineffective blocking variable in the ANOVA increases the chance of a Type II error in the primary hypothesis, H_0 . The conclusion of the primary hypothesis in this example would be more precise without the blocking variable. In fact, this is what essentially happened when we included the blocking variable with the randomized block design. With the blocking variable present in the model, we failed to discover a difference in the population means. Now go back to the beginning of the

chapter. When we tested the population means using one-way ANOVA (without a blocking variable), we concluded that the population means were indeed different.

In summary (It's about time!), if you feel there is a variable present in your model that could contribute undesirable variation, such as taking samples from different lawns, use the randomized block ANOVA. First test H_0 , the blocking hypothesis.

◆ If you reject H_0 , the blocking procedure was effective. Proceed to test H_m the primary hypothesis concerning the population means, and draw your conclusions.

◆ If you fail to reject H_0 , the blocking procedure was not effective. Redo the analysis using one-way ANOVA (without blocking) and draw your conclusions.

◆ If all else fails, take two aspirin and call me in the morning.

Your Turn

A consumer group is testing the gas mileage of three different models of cars. Several cars of each model were driven 500 miles and the mileage was recorded as follows.

306 Part k: Advanced Inferential Statistics

Car 1 Car 2 Car 3

22.5 18.7 17.2 20.8 19.8 18.0 22.0 20.4 21.1

23.6 18.0 19.8 21.3 21.4 18.6 22.5 19.7

Note that the size of each sample does not have to be equal for ANOVA.

Test for a difference between sample means using $\alpha = 0.05$.

2. Perform a pairwise comparison test on the sample means from Problem 1.

3. A vice president would like to determine whether there is a difference between the average number of customers per day between four different stores using the following data.

Store 1 Store 2 Store 3 Store 4

Note that the size of each sample does not have to be equal for ANOVA.

Test for a difference between sample means using $\alpha = 0.05$.

4. A certain unnamed statistics author and his two sons played golf at four different courses with the following scores:

Dad Brian John

Using completely randomized block ANOVA, test for the difference of golf score means using $\alpha = 0.05$ and using the courses as the blocking variable.

The Least You Need to Know

- ◆ Analysis of variance, also known as ANOVA, compares the means of three or more populations.
- ◆ A factor in ANOVA describes the cause of the variation in the data. When only one factor is being considered, the procedure is known as one-way ANOVA.
- ◆ A level in ANOVA describes the number of categories within the factor of interest.
- ◆ The simplest type of ANOVA is known as completely randomized one-way ANOVA, which involves an independent random selection of observations for each level of one factor.
- ◆ Completely randomized block ANOVA controls for variations from other sources than the factors of interest. This is accomplished by grouping the samples using a blocking variable.
- ◆ After rejecting H_0 using ANOVA, we can determine which of the

sample means are different using the Scheffe test.

Chapte

20

Correlation and Simple Regression

In This Chapter

- ◆ Distinguishing between independent and dependent variables
- ◆ Determining the correlation and regression line for a set of ordered-pair data
- ◆ Calculating a confidence interval for a regression line
- ◆ Performing a hypothesis test on the regression line
- ◆ Using Excel to perform simple regression analysis

For the last several chapters, we have put inferential statistics to work drawing conclusions about one, two, or more population means and proportions. I know this has been a lot of fun for you, but it's time to move to another type of inferential statistics that's even more exciting. (If you can imagine that!)

This final chapter focuses on describing how two variables relate to one another. Using correlation and simple regression, we will be able to first determine whether a relationship does indeed exist between the variables and second describe the nature of this relationship in mathematical terms. And hopefully we'll have some fun doing it!

310 Part k: Advanced Inferential Statistics

Independent Versus Dependent Variables

Suppose I would like to investigate the relationship between the number of hours that a student studies for a statistics exam and the grade for that exam (uh-oh). The following table shows sample data from six students whom I randomly chose.

Data for Statistics Exam

Hours Studied

Exam Grade

86 95 92

83 78 82

4ef

•rrrticio

The independent variable (x)

causes variation in the dependent variable (y).

*

\y^, Wrong Number

w'^P Exercise caution when deciding which variable is independent and which is dependent. Examine the relationship from both directions to see which one makes the most sense. The wrong choice will lead to meaningless results.

Obviously, we would expect the number of hours studying to affect the grade. The Hours Studied variable is considered the independent variable (x) because it causes the observed variation in the Exam Grade, which is considered the dependent variable (y). The data from the previous table is considered ordered pairs of (x,y) values, such as (3,86) and (5,95).

This "causal relationship" between independent and dependent variables only exists in one direction, as shown here:

Independent variable (x) —» Dependent variable (y)

This relationship does not work in reverse. For instance, we would not expect that the exam grade variable would cause the student to study a certain number of hours in our previous example.

Other examples of independent and dependent variables are shown in the following table.

Examples of Independent and Dependent Variables

Independent Variable

Dependent Variable

Size of TV

Level of advertising

Size of sports team payroll

Selling price of TV Volume of sales Number of games won

Now, let's focus on describing the relationship between the x and y variables using inferential statistics.

Correlation

Correlation measures both the strength and direction of the relationship between x and y. Figure 20.1 illustrates the different types of correlation in a series of scatter plots, which graphs each ordered pair of (x,y). The convention is to place the x variable on the horizontal axis and the y variable on the vertical axis.

* * Figure 20.1

Different types of correlation.

(A) Positive Linear Correlation

y

(B) Negative Linear Correlation

y

{C} No Correlation

(D) Nonlinear Correlation

Graph A in Figure 20.1 shows an example of positive linear correlation where, as x increases, y also tends to increase in a linear (straight line) fashion. Graph B shows a negative linear correlation where, as x increases, y tends to decrease linearly. Graph C indicates no correlation between x and y . This set of variables appears to have no impact on each other. And finally, Graph D is an example of a nonlinear relationship

312 Part 4: Advanced Inferential Statistics

between variables. As x increases, y decreases at first and then changes direction and increases.

For the remainder of this chapter, we will focus on linear relationships between the independent and dependent variables. Nonlinear relationships can be very disagreeable and go beyond the scope of this book.

Correlation Coefficient

The correlation coefficient, r , provides us with both the strength and direction of the relationship between the independent and dependent variables. Values of r range between -1.0 and $+1.0$. When r is positive, the relationship between x and y is positive (Graph A from Figure 20.1), and when r is negative, the relationship is negative (Graph B). A correlation coefficient close to 0 is evidence that there is no relationship between x and y (Graph C).

The strength of the relationship between x and y is measured by how close the correlation coefficient is to $+1.0$ or -1.0 and can be viewed in Figure 20.2.

Figure 20.2

The strength of the relationship.

The Strength of the Relationship

y

Graph A illustrates a perfect positive correlation between x and y with $r = +1.0$. Graph B shows a perfect negative correlation between x and y with $r = -1.0$. Graphs C and D are examples of weaker relationships between the independent and dependent variables.

We can calculate the actual correlation coefficient using the following equation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

y

def

•i # ni*tien

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The correlation coefficient, r , indicates both the strength and direction of the relationship between the independent and dependent variables.

Values of r range from -1.0 , a strong negative relationship, to $+1.0$, a strong positive relationship. When $r = 0$, there is no relationship between variables x and y .

Wow! I know this looks overwhelming, but before we panic, let's try out our exam grade example on this. The following table will help break down the calculations and make them more manageable.

JH Part 4: Advanced Inferential Statistics

Using these values along with $n = 6$, the number of ordered pairs, we have:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\frac{6(1835)-(21)(516)}{174} \cdot \frac{[6(79)-(21)^2][6(44582)-(516)^2]}{174}$$

174

$$V(33)(1236)$$

$$= 0.862$$

As you can see, we have a fairly strong positive correlation between hours of study and the exam grade. That's good news for us teachers.

*

Vv, Wrong Number

f ^ Be careful to distinguish between 2^x and (X^2) ■ With 2^x , we first square each value of x and then add each squared term. With (X^2) , we first add each value of x and then square this result. The answers between the two are very different!

What is the benefit of establishing a relationship between two variables such as these? That's an excellent question. When we discover that a relationship does exist, we can predict exam scores based on a particular number of hours of study. Simply put, the stronger the relationship, the more accurate our prediction will be. You will learn how to make such predictions later in this chapter when we discuss simple regression.

Testing the Significance of the Correlation Coefficient

We can perform a hypothesis test to determine whether the population correlation coefficient, ρ , is significantly different from 0 based on the value of the calculated correlation coefficient, r . We can state the hypotheses as:

$$H_0: \rho \leq 0$$

$$H_a: \rho > 0$$

This statement tests whether a positive correlation exists between x and y . I could also choose a two-tail test that would investigate whether any correlation exists (either positive or negative) by setting $H_0: \rho = 0$ and $H_a: \rho \neq 0$.

The test statistic for the correlation coefficient uses the Student's t -distribution as follows:

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ where:

r = the calculated correlation coefficient from the ordered pairs

n = the number of ordered pairs

For the exam grade example, the calculated t -statistic becomes: $t = 0.862$

The critical t -statistic is based on d.f. = $n - 2$ if we choose $\alpha = 0.05$, $t_{\alpha} = 2.132$ from Table 4 in Appendix B for a one-tail test. Because $t < t_{\alpha}$, we reject H_0 and conclude that there is indeed a positive correlation coefficient between hours of study and the exam grade. Once again, statistics has proven that all is right in the world!

Using Excel to Calculate Correlation Coefficients

After looking at the nasty calculations involved for the correlation coefficient, I'm sure you'll be relieved to know that Excel will do the work for you with the CORREL function that has the following characteristics:

CORREL(array1, array2)

where:

array1 = the range of data for the first variable
array2 = the range of data for the second variable

316 Part k: Advanced Inferential Statistics

For instance, Figure 20.3 shows the CORREL function being used to calculate the correlation coefficient for the exam grade example.

Figure 20.3

CORREL function in Excel with the exam grade example.

Cell C1 contains the Excel formula =CORREL(A2:A7,B2:B7) with the result being 0.862.

Simple Regression

The technique of simple regression enables us to describe a straight line that best fits a series of ordered pairs (x,y). The equation for a straight line, known as a linear equation, takes the form:

$$y = a + bx$$

where:

defini'tien

The technique of simple regression enables us to describe a straight line that best fits a series of ordered pairs [x,y).

y - the predicted value of y, given a value of x

x = the independent variable

a = the y-intercept for the straight line

b = the slope of the straight line

Figure 20.4 illustrates this concept.

Figure 20.4 shows a line described by the equation $y = 2 + 0.5x$. The y-intercept is the point where the line crosses the y-axis, which in this case is a - 2. The slope of the line, b, is shown as the ratio of the rise of the line over the run of the line, shown as $\frac{\text{rise}}{\text{run}} = 0.5$. A positive slope indicates the line is rising from left to right. A negative slope, you guessed it, moves lower from left to right. If $b = 0$, the line is horizontal, which means there is no relationship between the independent and dependent variables. In other words, a change in the value of x has no effect on the value of y.

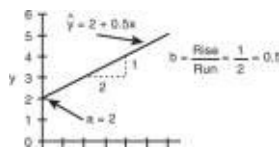


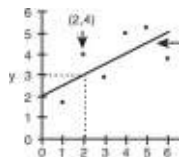
Figure 20.4

Equation for a straight line.

- \ -I-I-I-I-h

0 1 2 3 4 5 6

Students sometimes struggle with the distinction between y and y-hat. Figure 20.5 shows six ordered pairs and a line that appears to fit the data described by the equation $\hat{y} = 2 + 0.5x$.



- $\hat{y} = 2 + 0.5x$

$x = 2$

$y - \hat{y} = 3$

Figure 20.5

The difference between \hat{y} and y .

Figure 20.5 shows a data point that corresponds to the ordered pair $x = 2$ and $j = 4$. Notice that the predicted value of y according to the line at $x = 2$ is $y = 3$. We can verify this using the equation as follows:

$$\hat{y}_j = 2 + 0.5x = 2 + 0.5(2) = 3$$

The value of y represents an actual data point, while the value of \hat{y} is the predicted value of y using the linear equation, given a value for x .

Our next step is to find the linear equation that best fits a set of ordered pairs.

The Least Squares Method

The least squares method is a mathematical procedure to identify the linear equation that best fits a set of ordered pairs by finding values for a , the y -intercept; and b , the slope. The goal of the least squares method is to minimize the total squared error between the values of y and \hat{y} . If we define the error as $y - \hat{y}$ for each data point, the least squares method will minimize:

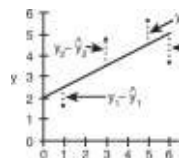
where n is the number of ordered pairs around the line that best fits the data.

318 Part 4: Advanced Inferential Statistics

This concept is illustrated in Figure 20.6.

Figure 20.6

Minimizing the error.



$y - \hat{y}$

0 1 2 3 4 5

According to Figure 20.6, the line that best fits the data, the regression line, will minimize the total squared error of the four data points. I'll demonstrate how to determine this regression equation using the least squares method through the following example.

def

•rni^tien

The least squares method is

a mathematical procedure to identify the linear equation that best fits a set of ordered pairs by finding values for a , the y-intercept; and b , the slope. The goal of the least squares method is to minimize the total squared error between the values of y and \hat{y} . The regression line is the line that best fits the data.

never to step foot into because me—you don't ever want to go

Apparently, there has been a silent war raging in our bathroom at home that has recently caught my attention. I'm of course referring to the space on our bathroom countertop that, under an unwritten agreement, Debbie and I are supposed to "share." Over the past few months, I have been keeping a wary eye on the increasing number of "things" on her side that are growing in number at a rate faster than the federal deficit. I'm slowly being squeezed out of my end of the bathroom by containers with words such as "volu-mizing fixative" and "soyagen complex." Debbie's little army has even taken both electrical outlets, cutting me off from any source of power. I might as well just raise the white towel in surrender and head off to the teenagers' bathroom in exile, a room I had vowed

... well, I'll just spare you the messy details. Just believe

in there voluntarily.

Anyway, let's say the following table shows the number of Debbie's items on the bathroom counter for the past several months.

Because my goal is to investigate whether the number of items is increasing over time, Month will be the independent variable and

Number of Items will be the dependent variable.

The least squares method finds the linear equation that best fits the data by determining the value for a, the y-intercept; and b, the slope, using the following equations:

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

where:

\bar{x} - the average value of x, the independent variable \bar{y} - the average value of y, the dependent variable

The following table summarizes the calculations necessary for these equations.

Calculations for the Slope and Intercept

320 Part 4: Advanced Inferential Statistics

$$a = \bar{y} - b\bar{x} = 10.5 - (0.976)5.5 = 5.13$$

The regression line for the bathroom counter example would be:

$$\hat{y} = 5.13 + 0.976x$$

Because the slope of this equation is a positive 0.976, I have evidence that the number of items on the counter is increasing over time at an average rate of nearly one per month. Figure 20.7 shows the regression line with the ordered pairs.

My prediction for the number of items on the counter in another six months (Month 16 from my data) will be:

$$\hat{y} = 5.13 + 0.976x = 5.13 + 0.976(16) = 20.7 = 21 \text{ items}$$

Look out, kids.
Make room for Dad.

15 in 11°

5

$$y = 5.13 + 0.976x$$

-1 h

-1

10 12

Figure 20.7

Regression line for the bathroom, counter example.

Confidence Interval for the Regression Line

Just how accurate is my estimate for the number of items on the counter for a particular month? To answer this, we need to determine the standard error of the estimate, s_r , using the following formula:

$$s_r = \sqrt{\frac{1}{n-2} \left(\sum y^2 - a^2 \sum y - b^2 \sum xy \right)}$$

The standard error of the estimate measures the amount of dispersion of the observed data around the regression line. If the data points are very close to the line, the standard error of the estimate is relatively low and vice versa. For our bathroom example:

$$s_r = \sqrt{\frac{1}{12-2} \left(1221 - 5.13(105) - 0.976(658) \right)}$$

10-2

40.14

8

:2.24

We are now ready to calculate a confidence interval (remember those from Chapter 14?) for the mean of $j <$ around a particular value of x . For

Month 8 ($x = 8$) in the data, Debbie has 11 items ($y = 11$) on the counter. The regression line predicted she would have:

$$y = 5.13 + 0.976x = 5.13 + 0.976(8) = 12.9 \text{ items}$$

definition

The standard error of the estimate, s_e , measures the amount of dispersion of the observed data around the regression line.

322 Part 4: Advanced Inferential Statistics

In general, the confidence interval around the mean of y given a specific value of x can be found by:

$$CI = \bar{y} \pm t$$

$$[x - x]$$

where:

t = the critical t-statistic from the Students' t-distribution
 s_e = the standard error of the mean
 n = the number of ordered pairs

Hold on to your hat while we dive into this one with our example. Suppose we would like a 95 percent confidence interval around the mean of y for Month 8. To find our critical t-statistic, we look to Table 4 in Appendix B. This procedure has $n - 2 = 10 - 2 = 8$ degrees of freedom, resulting in $t = 2.306$ from Table 4 in Appendix B. Our confidence interval is then:

$$CI = \bar{y} \pm t s_e$$

$$[x - x]$$

$$CI = 12.9 \pm (2.306)(2.24) = 12.9 \pm 5.16$$

$$[11, 18.06]$$

(385)-

$$CI = 12.9 + (2.306)(2.24)(0.419) = 12.9 + 2.16 \quad CI = 10.74 \text{ and } 15.06$$

This interval is shown graphically on Figure 20.8.

Our 95 percent confidence interval for the number of items on the counter in Month 8 is between 10.74 and 15.06 items. Sounds like a very crowded countertop to me.

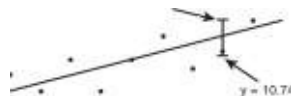
20

15

11 °

5

$y = 15.06$



-112

Figure 20.8

95 percent confidence interval for $x = 8$.

Testing the Slope of the Regression Line

Recall that if the slope of the regression line, b , is equal to 0, then there is no relationship between x and y . In our bathroom counter example, we found the slope of the regression line to be 0.976. However, because this result was based on a sample of observations, we need to test to see whether 0.976 is far enough away from 0 to claim a relationship really does exist between the variables. If β is the slope of the true population, then our hypotheses statement would be:

$$H_0: \beta = 0$$

$H_0: \rho = 0$

If we reject the null hypothesis, we conclude that a relationship does exist between the independent and dependent variables based on our sample. We'll test this using $\alpha = 0.01$.

This hypothesis test requires the standard error of the slope, s_b , which is found with the following equation:

, Wrong Number

$\frac{1}{n-2}$

$\sqrt{\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2}$ Just because a relationship

between two variables is statistically significant doesn't necessarily mean that a causal relationship truly exists. The mathematical relationship could be due to pure coincidence. Always use your best judgment when making these decisions.

where s_b is the standard error of the estimate that we calculated earlier.

For our bathroom example: s_b

2.24

$2.24 \sqrt{10}$

7.17

0.247

The test statistic for this hypothesis is:

$t = \frac{b_1}{s_b}$

where β_1 is the value of the population slope according to the null hypothesis.

For this example, our calculated t-statistic is:

$$t = \frac{0.976 - 0}{0.247} = 3.951$$

s, 0.247

The critical t-statistic is taken from the Student's t-distribution with $n - 2 = 10 - 2 = 8$ degrees of freedom. With a two-tail test and $\alpha = 0.01$, $t = 3.355$ according to Table 4 in Appendix B. Because $t > t_{\alpha/2}$, we reject the null hypothesis and conclude there is a relationship between the month and the number of items on the bathroom counter-top. I thought so!

The Coefficient of Determination

Another way of measuring the strength of a relationship is with the coefficient of determination, r^2 . This represents the percentage of the variation in y that is explained

by the regression line. We find this value by simply squaring r , the correlation coefficient. For the bathroom example, the correlation coefficient is:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

definition

The coefficient of determination, r^2 , represents the percentage of the variation in y that is explained by the regression line.

$$r^2 =$$

$$= \frac{(\sum (X - \bar{X})(Y - \bar{Y}))^2}{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}$$

$$= \frac{10(658) - (55)(105)}{10(658) - (55)(105)}$$

$$= \frac{[10(385) - (55)^2][10(1221) - (105)^2]}{10(658) - (55)(105)}$$

$$= \frac{V(825)(1185)}{10(658) - (55)(105)}$$

$$= 0.814$$

The coefficient of determination becomes: $r^2 = (.0814)^2 = 0.663$

In other words, 66.3 percent of the variation in the number of items on the counter is explained by the Month variable. If $r = 1$, all of the variation in y is explained by the variable x . If $r = 0$, none of the variation in y is explained by the variable x .

Using Excel for Simple Regression

Now that we have burned out our calculators with all these fancy equations, let me show you how Excel does it all for us.

1. Start by placing the bathroom counter data in Columns A and B in a blank sheet.
2. Go to the Tools menu and select Data Analysis. (Refer to the section "Installing the Data Analysis Add-in" from Chapter 2 if you don't see the Data Analysis command on the Tools menu.)
3. From the Data Analysis dialog box, select Regression as shown in Figure 20.9 and click OK.
4. Set up the Regression dialog box according to Figure 20.10.



il

8

imt |

Figure 20.10

Tfo Regression dialog box.

5. Click OK. Figure 20.11 shows the final regression results.

Figure20.il

Final results of the regression analysis if? Excel.

These results are consistent with what we found after grinding it out in the previous sections. Because the Rvalue for the independent variable Month is shown as 0.00414, which is less than $\alpha = 0.01$, we can reject the null hypothesis and conclude that a relationship between the variables does exist. Debbie has to believe me now!

A Simple Regression Example with Negative Correlation

Both of these past examples have involved a positive relationship between x and y . Now this example will summarize performing simple regression with a negative relationship.

Recently, I had the opportunity to "bond" with my son Brian as we shopped for his first car when he turned 16. Brian had visions of Mercedes and BMWs dancing in his head, whereas I was thinking more along the line of Hondas and Toyotas. After many "discussions" on the matter, we compromised on looking for 1999 Volkswagen Jettas. However, Brian had two requirements:

- ◆ It had to be black.
- ◆ It had to be the "new body" style.

Apparently, somebody at Volkswagen had the brilliant idea back in 1999 to subtly change the design of the Jetta halfway through the production year. Personally, I would never have noticed the difference. Brian, on the other hand, wouldn't be caught dead driving the original version, essentially eliminating half the used 1999 Volkswagen Jettas on the market. Undeterred, I searched the world over, asking each seller, "Is it the new body style?" Ahh, the joys of parenthood! Anyway, what follows

is a table showing the mileage of eight cars with the new body style and their asking price. The remainder of this chapter demonstrates the correlation and regression technique using this data.

Data for Car Example

Mileage Price Mileage Price

The following table, which shows the data in thousands, will be used for the various equations.

328 Part 4: Advanced Inferential Statistics

The correlation coefficient can be found using:

$$\frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{[\sum (x_i - \bar{x})^2] [\sum (y_i - \bar{y})^2]}}$$

$$\frac{8(5100.1) - (450.1)(96.2)}{\sqrt{[8(28786.8) - (450.1)^2] [8(1205.9) - (96.2)^2]}}$$

$$\frac{-2498.82}{(27794.4)(392.76)}$$

$$= -0.756$$

The negative correlation indicates that as mileage (x) increases, the price (y) decreases as we would expect. The coefficient of determination becomes:

$$r^2 = (-0.756)^2 = 0.572$$

Approximately 57 percent of the variation in price is explained by the variation in mileage. The regression line is determined using:

$$\bar{y} - b\bar{x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{8(5100.1) - (450.1)(96.2)}{8(28786.8) - (450.1)^2}$$

$$= \frac{-2498.82}{27794.4}$$

$$= -0.0902$$

$$= -0.0902$$

$$a = \bar{y} - b\bar{x} = 12.025 - (-0.0902)(56.26) = 17.100$$

We can describe the regression line by the equation: $y = 17.1 - 0.0902x$

This equation is shown graphically in Figure 20.12.

Figure 20.12

Recession line for car example.

Regression Line for Car Example

<: ".! .n■

115.000

110,000

JS.000

10

j^l?!.!o-°O.Wo2i

0 20.000 10.000 00.000 00,000 100.0m Mileage (xf

What would the predicted price be for a car with 45,000 miles? $y = 17.1 - 0.0902(45.0) = \$13,041$

The regression line would predict that a car with 45,000 miles would be priced at \$13,041. What would be the 90 percent confidence interval at $x = 45,000$? The standard error of the estimate would be:

$jY < y \ 2 \ a \ 1 < y_h \ Y < x \ y$

$(1205.9) - 17.1(96.2) - (-0.0902)(5100.1)$

$(1205.9) - (1645.02) + (460.03) , , , , s = , = 1.867$

$* \setminus 6$

The critical t-statistic for $n - 2 = 8 - 2 = 6$ degrees of freedom and a 90 percent confidence interval is $t_{\alpha/2} = 1.943$ from Table 4 in Appendix B. Our confidence interval is then:

$CI = \bar{y} \pm t_{\alpha/2} s_e$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

$\hat{y} = 13.041 + (1.934)(1.867)$

The 90 percent confidence interval for a car with 45,000 miles is between \$11,589 and \$14,493.

Is the relationship between mileage and price statistically significant at the $\alpha = 0.10$ level? Our hypotheses' statement is:

$H_0: \beta = 0$

$H_a: \beta \neq 0$

330 Part k: Advanced Inferential Statistics

The standard error of the slope, s_{b_1} , is found using: $s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$

$s_{b_1} = \frac{1.867}{\sqrt{28786.8}}$

$s_{b_1} = 0.0317$

$t_{\alpha/2, n-2} = 1.645$

The calculated test statistic for this hypothesis is: $t = \frac{b_1 - \beta_0}{s_{b_1}}$

$t = \frac{1.867}{0.0317}$

$t = 58.89$

The critical t-statistic is taken from the Student's t-distribution with 77 —

$2 = 8 - 2 = 6$ degrees of freedom. With a two-tail test and $\alpha = 0.10$ level, $t_{\alpha/2, df} = 1.943$ according to Table 4 in Appendix B. Because $|t| > t_{\alpha/2, df}$, we reject the null hypothesis and conclude there is a relationship between the mileage and price variable. We use the absolute values because the calculated t-statistic is in the left tail of the t-distribution with a two-tail hypothesis test.

Assumptions for Simple Regression

For all these results to be valid, we need to make sure that the underlying assumptions of simple regression are not violated. These assumptions are as follows:

- ◆ Individual differences between the data and the regression line, $(y_i - \hat{y}_i)$ are independent of one another.
- ◆ The observed values of y are normally distributed around the predicted value, \hat{y} .
- ◆ The variation of \hat{y} around the regression line is equal for all values of x .

Unfortunately (or fortunately), the techniques to test these assumptions go beyond the level of this book.

Simple Versus Multiple Regression

Simple regression is limited to examining the relationship between a dependent variable and only one independent variable. If more than one independent variable is involved in the relationship, then we need to graduate to multiple regression. The regression equation for this method looks like this:

y

$$= a + b_1x_1 + b_2x_2 + \dots + b_kx_k + \epsilon$$

As you can imagine, this technique gets really messy and goes beyond the scope of this book. I'll have to save this topic for *The Complete Idiot's Guide to Statistics, Part 3*. Uh-oh, I think I just heard Debbie faint.

Your Turn

1. The following table shows the payroll for 10 major league baseball teams (in millions) for the 2002 season, along with the number of wins for that year.

Payroll Wins Payroll Wins

Calculate the correlation coefficient. Test to see whether the correlation coefficient is not equal to 0 at the 0.05 level.

2. Using the data from Problem 1, answer the following questions:

a) What is the regression line that best fits the data?

b) Is the relationship between payroll and wins statistically significant at the 0.05 level?

c) What is the predicted number of wins with a \$70 million payroll?

d) What is the 99 percent confidence interval around the mean number of wins for a \$70 million payroll?

e) What percent of the variation in wins is explained by the payroll?

3. The following table shows the grade point average (GPA) for five students along with their entrance exam scores for MBA programs (GMAT). Develop a model that would predict the GPA of a student based on his GMAT score. What would be the predicted GPA for a student with a GMAT score of 600?

332 Part k: Advanced Inferential Statistics

Student GPA GMAT

The Least You Need to Know

◆ The independent variable (x) causes variation in the dependent variable (y).

◆ The correlation coefficient, r ; indicates both the strength and direction

of the relationship between the independent and dependent variables.

♦ The technique of simple regression enables us to describe a straight line that best fits a series of ordered pairs (x,y) .

♦ The least squares method is a mathematical procedure to identify the linear equation that best fits a set of ordered pairs by finding values for a , the y-intercept; and b , the slope.

♦ The standard error of the estimate, s_r , measures the amount of dispersion of the observed data around the regression line.

♦ The coefficient of determination, r^2 , represents the percentage of the variation in y that is explained by the regression line.

Append

Solutions to "Your Turn"



Chapter 1

1. Inferential statistics, because it would not be feasible to survey every Asian American household in the country. These results would be based on a sample of the population and used to make an inference on the entire population.

2. Inferential statistics, because it would not be feasible to survey every household in the country. These results would be based on a sample of the population and used to make an inference on the entire population.

3. Descriptive statistics, because Hank Aaron's home run total is based on the entire population, which is every at-bat in his career.

4. Descriptive statistics, because the average SAT score would be based on the entire population, which is the incoming freshman class.

5. Inferential statistics, because it would not be feasible to survey every American in the country. These results would be based on a sample of the population and used to make an inference on the entire population.

Chapter 2

1. Interval data, because temperature in degrees Fahrenheit does not contain a true zero point.

2. Ratio data, because monthly rainfall does have a true zero point.

3. Ordinal data, because a Master's degree is a higher level of education than a Bachelor's or high school degree. However, we cannot claim that a Master's degree is two or three times higher than the others.

4. Nominal data, because we cannot place the categories in any type of order.

5. Ratio data, because age does have a true zero point.

6. Definitely nominal data, unless you want to get into an argument about which is the lesser gender!

7. Interval data, because the difference between years is meaningful but a true zero point does not exist.

8. Nominal data, because I am not prepared to name one political party superior to another.

9. Nominal data, because these are simply unordered categories.

10. Ordinal data, because we can specify that "Above Expectation" is higher on the performance scale than the other two but we cannot comment on the differences between the categories.

11. Nominal data, because we cannot claim a person wearing the number "10" is any better than a person wearing the number "4."

12. Ordinal data, because we cannot comment on the difference in

performance between students. The top two students may be very far apart grade-wise, whereas the second and third students could be very close.

13. Ratio data, because these exam scores have a true zero point.

14. Nominal data, because there is no order in the states' categories.

Solutions to

335

Chapter 3

Exam Grade

Number of Students

56-60 61-65 66-70 71-75 76-80 81-85

5-90 91-95 96-100

Exam Scores

Histogram for Exam Grades

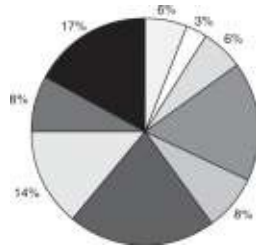
3.

Exam Grade

Number of Students

Percentage

Cumulative Percentage



Pie Chart for Exam Grades

5. 5|8 6 j 0268 7|222455899 8|1123455566689 9|125688999

Stem, and Leaf for Problem S

Solutions to "Your Turn" M7

6. 5(5)18 6 (0) j 02

6 (5) 168

7 (0) j 2224 7(5)|55899

8 (0)j 11234

8 (5) 155566689 9(0) j 12

9 (5) 15688999

Stem and Leaf for Problem 6

Chapter k

1. Mean = 15.9, Median = 17, Mode = 24

2. Mean = 81.7, Median = 82, Mode = 82

3. Mean = 32.7, Median = 32.5, Mode = 36 and 27

4. Mean = 7.2, Median = 6, Mode = 6

___ (22)(8) + (27)(37) + (32)(25) + (37)(48) + (42)(27) + (47)(10) _ 31

8 + 37 + 25 + 48 + 27 + 10 ' Y

$$6. ,J3)(118) + (2)(125) + (1)(107)^{\wedge} 1185$$

$$3 + 2 + 1 \quad 7 - _ (5)(1) + (7)(2) + (10)(3) + (8)(4) + (12)(5) + (3)(6) _ 3 \quad 5 + 7 \\ + 10 + 8 + 12 + 3$$

Chapter S

$$\pounds^* . =(120) \quad 2 = 14,400.$$

$$X^* \quad 2 \quad "$$

$$2,302-$$

$$14,400$$

$$n \quad -1$$

$$= 40.8,$$

$$y = V40.8 = 6.4, \text{ Range} = 24-8=16$$

$$\pounds^*=735 \quad \pounds^* \quad 2 = 60,207$$

$$= (735) \quad 2 = 540,225, \text{ a} \quad 2$$

$$X^* \quad 2 \quad "$$

$$Xr .$$

$$N$$

$$60,207$$

$$540,225$$

$$AT \quad 9$$

$$a = V20.2 = 4.5 , \text{ Range} = 90 - 75 = 15$$

$$3. \text{ Range} = 25, \text{ Variance} = 75.4, \text{ Standard Deviation} = 8.7$$

$$4. \text{ 2 5 6 6 6 8 10 11 11 15 } Q_1 = 5.5 \text{ } Q_2 = 7 \text{ } Q_3 = 11$$

Note the median of the data set is underlined.

:20.2

Solutions to "Your Turn" 339

$$(8)(22) + (37)(27) + (25)(32) + (48)(37) + (27)(42) + (10)(47) = 3158 + 37 + 25 + 48 + 27 + 10$$

$$f(x) = (x - \bar{x})^2 / (n - 1)$$

$$: \text{ £ } / = 155 \text{ £ } < > , - r)^2 / ; = 6,868.75$$

$$2 > / - *) 7,$$

$$\wedge \wedge = 744 \wedge 0 = 6.68 \text{ years } 155 - 1 \text{ y}$$

LUX*,

$$(5)(1) + (7)(2) + (10)(3) + (8)(4) + (12)(5) + (3)(6) = 35 + 5 + 7 + 10 + 8 + 12 + 3$$

$$f(x) = (x - \bar{x})^2 / (n - 1)$$

m

$$M_i = \frac{1}{n} \sum_{j=1}^n x_j$$

$$*'''''' r 22 \setminus = 1.49 \text{ years}$$

$$\setminus 77 - I \setminus 45 - I$$

7. Using the empirical rule, 95 percent of the values will fall within $k = 2$ standard deviations from the mean.

$$m. = k\sigma = 75 + 2(10) = 95, \setminus x = k\sigma = 75 - 2(10) = 55$$

Therefore, 95 percent of the data values should fall between 55 and 95.

8. The values 38 and 62 are two standard deviations from the mean of 50. This can be shown with the following:

$$|j. + ka = 62, |t + kc = 38$$

$$* = ^ = ^ = 2.0, , = -p^{\wedge}) = -pz50) = 2 . 0$$

Using Chebyshev's theorem, at least $1 - \frac{1}{k^2} \times 100\% = 1 - \frac{1}{2^2} \times 100\% = 75$ percent of the data values will fall between 38 and 62.

Chapter 6

la. Empirical, because we have historical data for Derek Jeter's batting average.

lb. Classical, because we know the number of cards and the number of aces in the deck.

lc. If I have data from my last several rounds of golf, this would be empirical, otherwise subjective.

Id. Classical, because we can calculate the probability based on the lottery rules.

le. Subjective, because I would not be collecting data for this experiment.

If. Subjective, because I would not be collecting data for this experiment.

Solutions to "Your Turn" lk\

2b. No, probability cannot be greater than 1.

2 c. No, probability cannot be greater than 100 percent.

2d. No, probability cannot be less than 1.

2e. Yes.

2f. Yes.

3a. $P[A] = \frac{1}{2} = 0.5$

3b. $P[B] = \frac{42}{125} = 0.336$

L' 125

3c. $P[A \text{ and } B] = P[A \cap B] = \frac{21}{125} = 0.168$

3d. The following table identifies the total number of families for the union of Events A and B.

Race Internet Number of Families

$P[A \text{ or } B] = P[B \cup A] = \frac{63}{125} = 0.504$

4a. $P[A \text{ and } B] = P[A \cap B] = \frac{20}{125} = 0.16$

4b. The following table identifies the total number of phone calls for the union of Events A and B.

Child Type of Call Number of Calls

$P[A \text{ or } B] = P[B \cup A] = \frac{64}{100} = 0.64$

Chapter 7

177

1. $P[A] = \frac{136}{260} = 0.523$

2 $P[B] = \frac{136}{260} = 0.523$

260

3 $P[A] = \frac{136}{260} = 0.523$

4 $P[B'] = 1 - P[B] = 1 - 0.523 = 0.477$

98

5. $P[A/B] = \frac{136}{260} = 0.523$

L J 152

$$6. P[A|B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{0.37}{0.58} = 0.64$$

152

79

$$7. P[A|B'] = \frac{P[A \text{ and } B']}{P[B']} = \frac{0.31}{0.42} = 0.73$$

108

$$8. P[A \text{ and } B] = P[A|B]P[B] = (0.64)(0.58) = 0.37$$

$$9. P[A \text{ and } B'] = P[A|B']P[B'] = (0.73)(0.42) = 0.31$$

$$10. P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B] = 0.68 + 0.58 - 0.37 = 0.89$$

$$11. P[A \text{ or } B'] = P[A] + P[B'] - P[A \text{ and } B'] = 0.68 + 0.42 - 0.31 = 0.79$$

$$12. P[B|A] =$$

$$\frac{P[B \text{ and } A] + P[B' \text{ and } A]}{P[A]}$$

$$= \frac{(0.58)(0.64) + 0.37}{0.68}$$

$$P[B|A] = \frac{0.58(0.64) + 0.37}{0.68} = 0.54$$

$$= \frac{(0.58)(0.64) + 0.37}{0.68} = 0.54$$

Chapter 8

$$1. 3 \times 8 \times 4 \times 3 = 288 \text{ different meals}$$

2. There are $4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 = 1,048,576$ different ways to answer the exam. If only one of these sequences is correct, the probability is $1/1,048,576 = 0.00000095$ that the student will correctly guess the correct sequence.

$$3. 13! = 6,227,020,800 \text{ different ordered arrangements}$$

Solutions to "Your Turn" 3H

1. Because $n = 10$, $r = 1$, $p = 0.5$

$$P[7, 10] = 10! (0.5)^7 (0.5)^{10-7} = 10! (0.5)^3 = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0.125 = 12.6$$

$$(10-7)! (0.5)^{10-7} = 3! (0.5)^3 = 6 \cdot 0.125 = 0.75$$

2. Because $n = 6$, $r = 3$, $p = 0.75$

$$P[3, 6] = \binom{6}{3} (0.75)^3 (0.25)^{6-3} = 20 \cdot (0.421875) = 8.4375$$

3. The probability of making at least 6 of 8 is $P[6, 8] + P[7, 8] + P[8, 8]$.
Because $y = 8$, $p = 0.8$

$$P[6, 8] = \binom{8}{6} (0.8)^6 (0.2)^{8-6} = 28 \cdot (0.262144) = 7.34$$

$$P[7, 8] = \binom{8}{7} (0.8)^7 (0.2)^{8-7} = 8 \cdot (0.2097152) = 1.6777216$$

$$P[8, 8] = \binom{8}{8} (0.8)^8 (0.2)^{8-8} = 1 \cdot (0.16777216) = 0.16777216$$

$$P[8, 8] = 1 \cdot (0.16777216) = 0.16777216$$

Therefore, the probability of making at least 6 out of 8 is $0.734 + 1.6777216 + 0.16777216 = 2.57949376$.

4. Because $n = 12$, $r = 6$, $p = 0.2$

$$P[6, 12] = \binom{12}{6} (0.2)^6 (0.8)^{12-6} = 924 \cdot (0.000064) = 0.0591456$$

5. The probability of no more than 2 out of the next 7 is $P[0, 7] + P[1, 7] + P[2, 7]$. Because $n = 7$, $p = 0.05$

$$P[0, 7] = \binom{7}{0} (0.05)^0 (0.95)^{7-0} = 1 \cdot (0.69834736875) = 0.69834736875$$

$$P[1, 7] = \binom{7}{1} (0.05)^1 (0.95)^{7-1} = 7 \cdot (0.05) \cdot (0.73509125) = 0.2573069375$$

$$P[2, 7] = \binom{7}{2} (0.05)^2 (0.95)^{7-2} = 21 \cdot (0.0025) \cdot (0.7737809375) = 0.040600000625$$

Therefore, the probability that no more than two of the next seven people will purchase is $0.69834736875 + 0.2573069375 + 0.040600000625 = 0.996254306875$.

Solutions to "Your Turn" 3W

6. Since $n = 4$, $p = 0.335$

$$P[0,4] = \frac{(4!)^2}{(4!)^2} (0.335)^0 (0.665)^4 = (1)(1)(0.196) = 0.196$$

$$P[1,4] = \frac{4!}{1!3!} (0.335)^1 (0.665)^3 = (4)(0.335)(0.294) = 0.394$$

$$P[2,4] = \frac{4!}{2!2!} (0.335)^2 (0.665)^2 = (6)(0.112)(0.442) = 0.297$$

$$P[3,4] = \frac{4!}{3!1!} (0.335)^3 (0.665)^1 = (4)(0.038)(0.665) = 0.101$$

$$P[4,4] = \frac{4!}{4!0!} (0.335)^4 (0.665)^0 = (1)(0.013)(1) = 0.013$$

$$P[4,0] = \frac{4!}{0!4!} (0.665)^4 (0.335)^0 = (1)(0.1296)(0.004096) = 0.1115$$

Chapter 10

$$i(6,4)(2.71838)^{-4} = (1296)(.002479) = 0.339$$

$$4! = 24$$

$$(7.55)(2.71838)^{-7.55} (23730.469)(0.000553)$$

$$2. P[5] = \frac{5!}{1!4!} = \frac{5!}{4!} = 0.1094$$

$$5! = 120$$

$$3. P[x > 2] = 1 - P[x < 2] = 1 - (P[x = 0] + P[x = 1] + P[x = 2])$$

$$(4.2^0)(2.71838)^{-4.2} (1)(.0150)$$

$$P[0] = \frac{4!}{0!4!} = \frac{4!}{4!} = 0.150$$

$$0! = 1$$

$$(4.2^1)(2.71838)^{-4.2} (4.2)(.0150)$$

$$P[1] = \frac{4!}{1!3!} = \frac{4!}{3!} = 0.630$$

$$1! = 1$$

$$(4.2^2)(2.71838)^{-4.2} (17.64)(.0150) \quad P[2] = \frac{4!}{2!2!} = \frac{4!}{2!} = 0.1323$$

$$P[x > 2] = 1 - (0.0150 + 0.0630 + 0.1323) = 0.7897$$

$$4. P[x < 3] = (P[x = 0] + P[x = 1] + P[x = 2] + P[x = 3])$$

$$(3.6^\circ)(2.7183836)(1)(.027324)$$

$$P[0] = \hat{\lambda} > S_1 = UV_1 = 0.0273$$

$$3.6 \cdot 1 \cdot 2.71838 - " 3.6 \cdot .027324$$

$$P[1] = i \wedge i = i - 'i > - = 0.0984$$

$$1! 1$$

$$(3.6 \cdot 3)(2.71838 - "') (12.96)(.027324)$$

$$p_2] = i ! S_1 = \setminus > \pm L = 0.1771$$

$$L J 2! 2$$

$$_ (3.6 \cdot 3)(2.71838 - \rangle) _ (46.656)(.027324) _$$

$$3! 6$$

$$P[\# < 3] = (0.0273 + 0.0984 + 0.1771 + 0.2125) = 0.5152$$

$$(2.5')(2.71838:s)(2.5)(.082085)$$

$$5. P[1] = \hat{\lambda} \wedge \pounds = * - \hat{i} = 0.2052$$

$$1! 1$$

$$6. P[x] = v ' rj , \rangle = 25, p = 0.05, \langle p = 1.25$$

$$I \rangle p) \bullet e$$

$$p[2] = a - 25) 2 \wedge ' = (1.5625)(0.2865 \ 0 \ 5) =$$

$$2! 2$$

$$(2.5^\circ)(2.7183825)(1)(.082085) \ 7. P[0] = \hat{\lambda} \wedge \pounds = \hat{\lambda} \wedge = 0.0821$$

Chapter 11

$$< 0 ' \blacksquare ^\circ 6 + 0.78, P[z > +0.78] = 1 - P[z < +0.78] = 1 - 0.7834 = 0.2166$$

$$-1.22, P[z > -1.22] = P[z < +1.22] = 0.$$

$$+2.0, P[z < +2.0] = 0.9772,$$

$$P[-0.43 < z < +2.0] = P[z < +2.0] - P[z < -0.43] = 0.9772 - 0.3327 = 0.6445$$

$$2a. Z_{90} = 1.90 \sim 1.76 = +0.63, P[z < +0.63] = 0.7349$$

$$2b. Z_{18} = 1.58 \sim 1.76 = -0.81, P[z < -0.81] = 1 - P[z < +0.81] = 0.2098$$

Solutions to "Your Turn" 3H

$$Z_{168} = 1.68 \sim 1.76 = -0.36, Z_{150} = 1.50 \sim 1.76 = -0.36$$

$$2c. Z = -0.36, Z_{150} = -1.17$$

$$P[-1.17 < z < -0.36] = P[z < -0.36] - P[z < -1.17]$$

$$P[z < -0.36] = 1 - P[z < +0.36] = 0.3599, P[z < -1.17] = 1 - P[z < +1.17] = 0.1218$$

$$P[-1.17 < z < -0.36] = 0.3599 - 0.1218 = 0.2381$$

$$3a. Z = 31 \sim 37.5 = -0.86, P[z > -0.86] = P[z < +0.86] = 0.8038$$

$$3b. Z_{7} = 42 \sim 37.5 = +0.59, P[z < +0.59] = 0.7231$$

$$7.45 - 37.5 = -0.45 \sim -0.59, Z_{40} = 40 \sim 37.5 = +0.33$$

$$3c. Z = +0.99, Z_{40} = +0.33$$

$$P[+0.99 < z < +0.33] = P[z < +0.99] - P[z < +0.33] = 0.8381 - 0.6289 = 0.2092$$

$$4. \text{ For this problem, } n = 14, p = 0.5, \text{ and } q = 0.5. \text{ We can use the normal approximation since } np = nq = (14)(0.5) = 7. \text{ The binomial probabilities}$$

from the binomial table are $P[r = 4, 5, \text{ or } 6] = 0.0611 + 0.1222 + 0.1833 = 0.3666$. Also, $H = np = (14)(0.5) = 7$ and $\sigma = \sqrt{npq} = \sqrt{(14)(0.5)(0.5)} = 1.871$. The normal approximation would be finding $P[3.5 < x < 6.5]$.

$$6.5 - 0.5 = 6, \quad 3.5 - 0.5 = 3$$

$$z_{.6} = -0.27, \quad z_{.3} = -1.87$$

$$6 - 1.871 = 4.129, \quad 3 - 1.871 = 2.129$$

$$P[-1.87 < z < -0.27] = P[z < -0.27] - P[z < -1.87]$$

$$P[z < -0.27] = 1 - P[z < +0.27] = 0.3946, \quad P[z < -1.87] = 1 - P[z < +1.87] = 0.0307$$

$$P[-1.87 < z < -0.27] = 0.3946 - 0.0307 = 0.3639$$

$$5a. z_{.97} = 1.92, \quad P[z > 1.25] = 1 - P[z < 1.25] = 1 - 0.8944 = 0.1056$$

$$5b. 2q = 0.50, \quad P[z > -0.50] = P[z < +0.50] = 0.6915$$

$$, 4000 - 4580 = -580, \quad 5000 - 4580 = 420$$

$$6a. z_{.4000} = -1.05, \quad z_{.5000} = 0.76$$

$$P[-1.05 < z < 0.76] = P[z < 0.76] - P[z < -1.05]$$

$$P[z < 0.76] = 0.7764, \quad P[z < -1.05] = 1 - P[z < +1.05] = 1 - 0.8531 = 0.1469$$

$$P[-1.05 < z < 0.76] = 0.7764 - 0.1469 = 0.6295$$

$$6b. . = 4200 \sim 4580 = -380, \quad P[z < -0.69] = 1 - P[z < +0.69] = 0.2421$$

$$4200 - 4580 = -380$$

Chapt

erl2

$$1. * = \wedge = \wedge = 150$$

n 500

2. If every employee belonged to a particular department, certain departments could be chosen for the survey, with every individual in those departments asked to participate. Other answers are also possible.

3. If each employee can be classified as either a manager or a nonmanager, ensure that the sample proportion for each type is similar to the proportion of managers and nonmanagers in the company. Other answers are also possible.

Chapter 13

1a. $\hat{p} = 0.58$

sh. V_{15}

, $i = 471$ «

1b. $O = 1.36$

1c. $c_7 = 1.57$ hi V_{20}

2a. $a = 2.5$, $z = +0.40$, $P[z < +0.40] = 0.6554$

* $V = 2.5$

2b. $z = 18 - 16 = +0.80$, $P[z > +0.80] = 1 - P[z < +0.80] = 1 - 0.7881 = 0.2119$

$14.5 - 16 = -1.5$, $z = -0.60$, $z = +0.20$

14.5 to 16.5

$P[z < +0.20] = 0.5793$, $P[z < -0.60] = 1 - P[z < +0.60] = 0.2743$ $P[14.5 < x < 16.5] = P[-0.60 < z < 0.20] = 0.5793 - 0.2743 = 0.3050$

Solutions to "Your Turn" JW

$p(1-p) = 0.25(1-0.25)$

$$3a < t, = \frac{1}{\sqrt{2\pi}} = \frac{1}{\sqrt{2}} = 0.0306$$

$$ia - \frac{1}{\sqrt{n}} \approx 200$$

$$p(1-p) = 0.42(1-0.42) = 0.2436$$

$$s^2 = 0.0494$$

$$it} - \frac{1}{\sqrt{n}} \approx 100$$

$$p(1-p) = 0.06(1-0.06) = 0.0564$$

$$p(1-p) = 0.32(1-0.32) = 0.2176$$

$$4a - \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{25}} = 0.2$$

$$P[z < -0.54] = 1 - P[z < +0.54] = 0.2946$$

$$4b. z = \frac{0.36 - 0.32}{\sqrt{0.0369}} = 1.08, P[z > +1.08] = 1 - P[z < +1.08] = 1 - 0.8599 = 0.1401$$

$$, 0.29 - 0.32 = -0.03$$

$$4c. z_{0.05} = -0.81, z_{0.95} = +1.36$$

$$0.29 - 0.32 = -0.03$$

$$P[z < +1.36] = 0.9131, P[z < -0.81] = 1 - P[z < +0.81] = 0.2090$$

$$e p(1-p) = 0.4(1-0.4) = 0.24$$

$$P[z > +1.51] = 1 - P[z < +1.51] = 0.0655$$

Chapter 10

$$1. \hat{\mu} = 1.20, z = 2.17$$

$$s^2 = 1.20$$

$$\text{Upper Limit} = \bar{x} + z \cdot \frac{s}{\sqrt{n}} = 31.3 + 2.17(1.20) = 33.90$$

$$=31.3-2.17(1.20) = 28.70$$

$$2. \text{ , } = (fJ=[^{|M}) 2 =48.9 \gg 49$$

3. This is a trick question! The sample size is too small to be used from a population that is not normally distributed. This question goes beyond the scope of this book. You would need to consult a statistician.

4. Using Excel, we can calculate $\hat{x} = 13.9$ and $\hat{s} = 6.04$.

$$s = 6.04 \cdot \ln a = -; = -j^{\wedge} = = 1.10$$

$$* 4n V_{30}$$

$$\text{Upper Limit} = x + z c d = 13.9 + 1.64(1.10) = 15.70 \quad \text{Lower Limit} = x - z r d = 13.9 - 1.64(1.10) = 12.10$$

5. Using Excel, we can calculate $\hat{x} = 46.92$.

$$<7 = 12.7, z = 1.88, \text{ and } a = -\% = -f = 3.67$$

$$* 4n V_{12}$$

$$\text{Upper Limit} = x + z t o = 46.92 + 1.88(3.67) = 53.82$$

$$\text{Lower Limit} = \sim x - z, a \sim = 46.92 - 1.88(3.67) = 40.02$$

$$s = 11.29$$

6. Using Excel, we can calculate $\bar{x} = 119.64$, $s = 11.29$, $a = -j = -p = - = 3.40$.

$$* V/2 v_{11}$$

For a 98 percent confidence interval with $n - 1 = 11 - 1 = 10$ degrees of freedom,

$$f. = 2.764.$$

$$\text{Upper Limit} = x + t c d = 119.64 + 2.764(3.40) = 129.04 \quad \text{Lower Limit} = x - t r d = 119.64 - 2.764(3.40) = 110.24$$

7. This is another trick question! The sample size is too small to be used from a population that is not normally distributed. This question goes beyond the scope of this book. You would need to consult a statistician.

8. $p_s = 0.055$. Since $np_s = (200)(0.055) = 11$ and $nq_s = (200)(0.945) = 189$,

we can use the normal approximation.

$$P, (1 - P,) 0.055) 0.945)$$

$$a = .M - ^ - = J - ^ = 0.0161, z f = 1.96$$

$$" V n \setminus 200$$

$$\text{Upper Limit} = p_t + z, a f = 0.055 + 1.96(0.0161) = 0.087 \quad \text{Lower Limit} = p_s - z e o = 0.055 - 1.96(0.0161) = 0.023$$

$$9. , = w [|J = (0.55)(0.45)(^J = 650$$

Solutions to "Your Turn" 351

Chapter IS

$$1. H_0 : \mu = 1.7, H_a : \mu \neq 1.7$$

$$n = 15, a = 0.5 \text{ cups}, o - = -j = -j = 0.0845 \text{ cups}, z c = \pm 1.64$$

$$\text{Upper limit} = (\mu H) + z p \sim = 1.7 + (1.64)(0.0845) = 1.84 \text{ cups}$$

$$\text{Lower limit} = \mu H - z c o - = 1.7 - (1.64)(0.0845) = 1.56 \text{ cups}$$

Since $x = 1.95$ cups, we reject H_0 and conclude that the population mean is not 1.7 cups per day.

$$2. H_0 : \mu \leq 40, H_a : \mu > 40, n = 50, a = 12.5 \text{ years}, a - = -? = \mu = 1.768 \text{ years}, z, = -1.64 \quad \wedge' 2 \wedge 50$$

$$\text{Lower limits} (\mu H + z c o - = 40 + (-1.64)(1.768) = 37.1 \text{ years}$$

Since $x = 38.7$ years, we do not reject H_0 and conclude that we do not

have enough evidence to support the claim that the average age is less than 40 years old.

3. $H_0: (\mu < 1000)$, $H_a: (\mu > 1000)$, $n = 32$, $\sigma = 325$ hours,

$\bar{x} = 1190$ hours, $z = \frac{1190 - 1000}{325/\sqrt{32}} = 2.05$

Upper limit = $1000 + (2.05)(325) = 1117.8$ hours

Since $\bar{x} = 1190$ hours, we reject H_0 and conclude the average light bulb life exceeds 1000 hours.

4. $H_0: (\mu \geq 30)$, $H_a: (\mu < 30)$, $n = 42$, $\sigma = 8.0$ minutes, $\bar{x} = 26.9$ minutes,

$z = \frac{26.9 - 30}{8.0/\sqrt{42}} = -2.33$

Lower limit = $30 + (-2.33)(8.0) = 27.13$ minutes

Since $\bar{x} = 26.9$ minutes, we reject H_0 and conclude that the average delivery time is less than 30 minutes.

5. $H_0: (\mu = \$2700)$, $H_a: (\mu \neq \$2700)$, $n = 40$, $\sigma = \$150.20$, $\bar{x} = \$2450$, $z = \frac{2450 - 2700}{150.20/\sqrt{40}} = -2.58$

Upper limit = $2700 + (1.96)(150.20) = \2994

Lower limit = $2700 - (1.96)(150.20) = \2406

Since $\bar{x} = \$2450$, we do not reject H_0 and conclude that we do not have enough evidence to contradict the claim that the average college student has \$2,700 in credit card debt.

Chapt

er16

1. $H_0: (\mu = 1100)$, $H_a: (\mu \neq 1100)$, $n = 70$, $\sigma = 310$, $\bar{x} = 1035$, $z = \frac{1035 - 1100}{310/\sqrt{70}} = -1.71$

$\alpha = 0.05$

$\alpha/2 = 0.025$, $z_{\alpha/2} = 1.96$

$$z = \pm 1.64, z = - = -1.75,$$

$$\alpha = 0.05$$

$$p\text{-value} = (2)(P[z < -1.75]) = (2)(1 - P[z < +1.75]) = (2)(1 - 0.9599) = 0.0802,$$

Since $p\text{-value} < \alpha$, we reject H_0 and conclude that the average SAT score does not equal 1100.

$$2. H_0: \mu = 35, H_a: \mu \neq 35, \bar{x} = 37.9, s = 6.74, b = 10, d.f. = n - 1 = 9,$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{37.9 - 35}{6.74/\sqrt{10}} = 1.36$$

$$(t_{\alpha/2, df} = t_{0.025, 9} = 1.833, t = 1.36 < 1.833, \text{ do not reject } H_0)$$

$$V \gg V_0 \text{ at } 2.13$$

Since $t < t_{\alpha/2}$, we do not reject H_0 and conclude that the average class size equals 35 students.

$$3. H_0: \mu \leq 7, H_a: \mu > 7, \bar{x} = 8.2, s = 4.29, n = 30, \alpha = 0.05, t_{\alpha, df} = t_{0.05, 29} = 1.699, t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{8.2 - 7}{4.29/\sqrt{30}} = 1.54$$

$$V \ll v_{30}$$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{8.2 - 7}{4.29/\sqrt{30}} = 1.54, z = 1.54 < z_{\alpha} = 1.64, \text{ do not reject } H_0$$

$$p\text{-value} = P[z > 1.54] = 1 - P[z < 1.54] = 1 - 0.9382 = 0.0618$$

Since $z < z_{\alpha}$ or $p\text{-value} > \alpha$, we do not reject H_0 and conclude that average gasoline consumption in the U.S. does not exceed seven liters per car per day.

$$\sim P - P_n \text{ at } 0.30 - 0.40, \dots$$

$$z = L = -2.70$$

$$\alpha = 0.037$$

$$p\text{-value} = P[z < -2.70] = 1 - P[z < +2.70] = 1 - 0.9965 = 0.0035, z_{\alpha} = -2.33$$

Since p-value < α , we reject H_0 and conclude that the proportion of Republicans is less than 40 percent.

Solutions to "Your Turn" 353

5. $H_0: p = 0.65$, $H_1: p \neq 0.65$, $\alpha = 0.05$, $n = 100$, $SS = 0.032$,

$p = 0.69 - 0.65 = 0.04$,

$z = \frac{p - p_0}{\sqrt{p_0(1-p_0)}} = \frac{0.04}{\sqrt{0.65(1-0.65)}} = +1.25$.

From a p 0.032

p-value = $(2)P[z > +1.25] = (2)(1 - P[z < +1.25]) = (2)(1 - 0.8944) = 0.2122$,

Since p-value > α , we fail to reject H_0 and conclude that the proportion of teenagers who exceed their minutes equals 65 percent.

6. $H_0: \mu \geq 15$, $H_1: \mu < 15$, $n = 60$, $\alpha = 0.05$, $cr_{\alpha} = -z_{\alpha} = -1.645$

$\bar{x} = 13.5$

$s = 3.5$

$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{13.5 - 15}{3.5/\sqrt{60}} = -2.33$.

$\alpha = 0.05$

p-value = $P[z < -2.33] = (1 - P[z < +2.33]) = (1 - 0.99) = 0.010$,

Since p-value < α , we reject H_0 and conclude that the average number of hours worked is less than 15.

Chapter 17

1. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $\alpha = 0.05$, $n_1 = 45$, $n_2 = 38$

$\bar{x}_1 = 24.22$, $\bar{x}_2 = 24.22$, $s_1 = 1.96$, $s_2 = 1.96$

$(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) = 0$, $SS_1 = 1.96^2(45-1) = 171.36$, $SS_2 = 1.96^2(38-1) = 147.36$

ff io 24.22

Since $z < z_{\alpha}$, we do not reject H_0 and conclude there is not enough evidence to support a difference between the two states.

$$p\text{-value} = (2)P[z > +0.91] = (2)(1 - P[z < +0.91]) = 2(1 - 0.8186) = 0.3628$$

$$2. H_0: \mu = \mu_0, H_a: \mu \neq \mu_0, \bar{x} = 88.3, s = 7.30, n = 82.4, \alpha = 0.05,$$

$$K-1, f + (n-1) = 82.4 - 1 = 81.4, H(81.4)(7.30)^2 + (9)(6.74)^2 = 701$$

$$11 + 10 = 21$$

$$a - s = t - (7.04) / (7.04) \sqrt{0.1909} = 3.08$$

$$\{X_1, X_2, \dots, X_n\} \sim (\mu, \sigma^2) \quad (88.3 - 82.4) / 0$$

$$\{ = ' = +1.92$$

$$cr - 3.08$$

$$d.f. = n - 2 = 11 + 10 - 2 = 19, t = +1.729$$

Since $f > t_{\alpha}$, we reject H_0 and conclude the satisfaction scores are not equal between the two stores.

$$3. H_0: \mu \leq 15, H_a: \mu > 15, \sum x = 21 + 23 + 11 + 19 + 15 + 20 + 17 + 23 + 17 = 166$$

$$\$V = 44! + 529 + 121 + 361 + 225 + 400 + 289 + 529 + 289 = 3184$$

$$-y^{166}$$

$$h_9$$

$$E, -(M - J(184, -W!))^{2222}$$

$$b^{-1} \setminus 8V8$$

$$t = \hat{\mu} = \hat{Z} = M_4 = +2.64, \sigma = B - 1 = 9 - 1 = 8, \alpha = +1.860, j > 3^{91} 1.30$$

J

■Jn V9

Since $t > t_r$, we reject H_0 and conclude the weight loss program claim is valid.

4. $H_0: p_1 \leq p_2$, $H_1: p_1 > p_2$, Pop 1 = Florida, Pop 2 = Nation

$n_1 = 272$, $n_2 = 390$, $x_1 = 157$, $x_2 = 252$

$\hat{p}_1 = \frac{157}{272} = 0.577$, $\hat{p}_2 = \frac{252}{390} = 0.646$, $p = 0.662$

$n_1 = 272$, $n_2 = 390$, $x_1 = 157$, $x_2 = 252$

$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.577 - 0.646}{\sqrt{0.662(1-0.662) \left(\frac{1}{272} + \frac{1}{390} \right)}} = -0.98$

$\alpha = 0.05$

$z_{\alpha} = 1.645$

Since $-0.98 > 1.645$, we do not reject H_0 and conclude there is not enough evidence to support the claim that the proportion of home ownership in Florida is greater than the national proportion.

p-value = $P[z > -0.98] = 1 - P[z < -0.98] = 1 - 0.8365 = 0.1635$

Solutions to "Your Turn" 355

5. $H_0: \mu_1 \leq \mu_2$, $H_1: \mu_1 > \mu_2$, Pop 1 = City A, Pop 2 = City B.

$\bar{x}_1 = 9.8$, $\bar{x}_2 = 9.1$, $s_1^2 = 0.4$, $s_2^2 = 0.8$

$n_1 = 10$, $n_2 = 10$

$t = 1.67$

$\alpha = 0.05$

p-value = $P[t > 1.67] = 1 - P[t < 1.67] = 1 - 0.9525 = 0.0475$

Since p-value < α , we reject H_0 and conclude that the average wage in

City A is higher than City B.

$$6. H_0: \mu_A = \mu_B, H_1: \mu_A \neq \mu_B, \bar{x}_A = 14.58, \bar{x}_B = 17.45, s_A^2 = 8.26$$

$$s_B^2 = 7.2, n_A = 12, n_B = 11$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} = \frac{4.19 - 6.20}{\sqrt{5.447 \left(\frac{1}{12} + \frac{1}{11} \right)}} = -0.890, t_{\alpha/2, df} = 1.725$$

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{11(8.26) + 10(7.2)}{21} = 7.69$$

$$s_p = \sqrt{7.69} = 2.77$$

$$t = \frac{4.19 - 6.20}{\sqrt{5.447 \left(\frac{1}{12} + \frac{1}{11} \right)}} = -0.890$$

$$t_{\alpha/2, df} = 1.725$$

$$|t| = 0.890 < 1.725$$

$$H_0: \mu_A = \mu_B, H_1: \mu_A \neq \mu_B$$

$$t = -0.890, t_{\alpha/2, df} = 1.725$$

Since $|t| < t_{\alpha/2, df}$, we fail to reject H_0 and conclude that the number of days that a home is on the market in City A does not differ from City B.

Chapter 18

1. H_0 : The arrival process can be described by the expected distribution
 H_1 : The arrival process differs from the expected distribution

For $\alpha = 0.05$ and d.f. = $k - 1 = 6 - 1 = 5$, $\chi^2_{\alpha/2} = 11.070$. Since $\chi^2 < \chi^2_{\alpha/2}$, we do not reject H_0 and conclude that the arrival distribution is consistent with the expected distribution.

H_0 : The process can be described with the Poisson distribution using $\lambda = 3$.
 H_1 : The process differs from the Poisson distribution using $\lambda = 3$.

Solutions to"

Number of Hits Per Minute

Poisson Probabilities

Number of Hits

Expected Frequency

Hits

(O-E)

For $\alpha = 0.01$ and d.f. = $1 - 1 = 8 - 1 = 7$, $\chi^2 = 18.475$. Since $\chi^2 > \chi^2_{\alpha}$, we do not reject H_0 and conclude that the process is consistent with the Poisson distribution using $\lambda = 3$.

3. H_0 : Grades are independent of reading time H_1 : Grades are dependent of reading time

Sample expected frequency calculations:

$(265)(95)/$

500

50.35 E_{11} ,

$(265)(128)/500$

$= 67.84$

$(265)(155)/$

$u 500$

(O-E)

For $\alpha = 0.05$ and d.f. = $(r - 1)(c - 1) = (3 - 1)(5 - 1) = 8$, $\chi^2_{\alpha} = 15.507$. Since $\chi^2 > \chi^2_{\alpha}$, we reject H_0 and conclude that there is a relationship between grades and the number of hours reading.

4. H_0 : The process can be described with the Binomial distribution using $p = 0.4$. H_1 : The process differs from the Binomial distribution using $p = 0.4$.

Solutions to"

359

For $\alpha = 0.05$ and d.f. = $k - 1 = 6 - 1 = 5$, $\chi^2_{0.05, 5} = 11.070$. Since $\chi^2_{calc} < \chi^2_{crit}$, we reject H_0 and conclude that the process differs from the Binomial distribution using $p = 0.4$.

Chapter 19

1. $H_0 : \sigma^2 = \sigma_0^2 = 22.12$

H_a : not all σ_i^2 's are equal $\alpha = 0.05$

$V = 17.77, n = 6$

$SSB = \sum_{j=1}^k n_j(\bar{x}_j - \bar{x})^2 = (6-1)0.98^2 + (6-1)1.45^2 + (5-1)2.36^2 = 21.59 = \frac{1}{n} \sum_{j=1}^k n_j \bar{x}_j^2 - \frac{(\sum_{j=1}^k n_j \bar{x}_j)^2}{n}$

N

17

A'

$SSB = \sum_{j=1}^k n_j(\bar{x}_j - \bar{x})^2 = 6(22.12 - 20.32)^2 + 6(19.67 - 20.32)^2 + 5(18.94 - 20.32)^2 = 31.50$ $MSB = \frac{SSB}{k-1} = \frac{31.50}{5} = 6.30$ $SSW = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = 21.59$ $MSW = \frac{SSW}{n-k} = \frac{21.59}{11} = 1.96$

$F_{0.05, 5, 11} = 3.98$

$F_{calc} = 3.17 < 3.98$

$F_{calc} < F_{crit}$

$MSW = 1.54$ $\sigma^2 \sim 1.54$

Since $F_{calc} < F_{crit}$, we reject H_0 and conclude that there is a difference between

the sample means.

2. For x_1 and x_2 , F

$$\{x_i - \bar{x}\}^2 = (22.12 - 19.67)^2 =$$

SSW

i

$- + -$

n

21.59

$$5 + 5 + 4$$

11

$$- + -66$$

70

For x , and s^2 , s

$$\{i - \bar{i}\}^2 = (22.12 - 18.94)^2 =$$

For x , and x , F

: 17.88

SSW

$i = \backslash$

$$F_{S, C} = (i-1) V_{u-i} = (3-1)(3.739) = 7.478$$

21.59

$$5 + 5 + 4$$

11

—+ -6 5

:0.94

Sample Pair

Conclusion

We conclude that there is a difference between gas mileage of Cars 1 and 2 and Cars 1 and 3.

Solutions to "Your Turn" 361

3. $H_0: \mu_1 = \mu_2 = \mu_3$ vs $H_1: \text{not all } \mu\text{'s are equal}$
 $\bar{x}_1 = 38.33$, $\bar{x}_2 = 28.29$, $\bar{x}_3 = 31.43$
 $s_1^2 = 115.47$, $s_2^2 = 12.51$, $s_3^2 = 86.8$
 $n_1 = 6$, $n_2 = 4$, $n_3 = 7$

$$SS_{\text{between}} = (6-1)115.47 + (4-1)12.51 + (7-1)86.8 = 2242.49$$

$$= Y^* 36 + 48 + 32 + 28 + \dots + 36 + 18 + 30 + 21, \dots, D = \bar{x} = 31.38$$

TV 26

$$SSB = 6(38.33 - 31.38)^2 + 4(28.29 - 31.38)^2 + 7(31.43 - 31.38)^2 = 425.22$$

$$MSB = SSB / (k-1) = 425.22 / 2 = 212.61$$

k-1 4-1 TV-£ 26-4

$$F = MSB / MSW = 212.61 / 101.93 = 2.09$$

Since $F < F_{\alpha}$, we do not reject H_0 and conclude that there is no difference between the sample means.

4. $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$
Pop 1 = Dad,

$H_1: \text{not all } \mu\text{'s are equal}$
Pop 2 = Brian,

TV = 12, k = 3, b = 4, Pop 3 = John,

H_0 : the block means are all equal

H_1 : the block means are not all equal

$$= Y_{.1} + Y_{.2} + Y_{.3} + Y_{.4} + \dots + Y_{.k} + Y_{.k+1} + Y_{.k+2} + Y_{.k+3}$$

$$\bar{y}_{..} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^b y_{ji} = 86.5$$

TV 12

$$(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$, j=1, 2, \dots, k$$

$$SSB = \sum_{j=1}^k n_j (\bar{y}_{.j} - \bar{y}_{..})^2 = 4(92.5 - 86.5)^2 + 4(83.5 - 86.5)^2 + 4(83.5 - 86.5)^2 = 216$$

$$SSBL = \sum_{j=1}^k (b-1) (\bar{y}_{.j} - \bar{y}_{..})^2 \text{ where } b = 1 \text{ for Course 1, etc.}$$

Block (course) averages: $\bar{y}_{.1} = 86$, $\bar{y}_{.2} = 91$, $\bar{y}_{.3} = 85$, $\bar{y}_{.4} = 84$

$$SSBL = (3)(86 - 86.5)^2 + (3)(91 - 86.5)^2 + (3)(85 - 86.5)^2 + (3)(84 - 86.5)^2 = 87$$

$$SSIV = SST - SSB - SSBL = 329 - 216 - 87 = 26$$

$$MSW = \frac{SSIV}{N - k - b + 1} = \frac{26}{12 - 3 - 4 + 1} = 4.33. \quad MSBL = \frac{SSBL}{k - 1} = \frac{87}{3} = 29, \quad (F_{3,8} = 2.75)$$

MSF 4.33

$$F_{3,8} = 2.75, \quad F_{3,8} = 2.75, \quad F_{3,8} = 2.75, \quad F_{3,8} = 2.75$$

Since $F > F_{\alpha}$, we reject H_0 and conclude that the blocking procedure was effective and proceed to test H_0 ■

$$MS_5 = \frac{SS_5}{N - k - b + 1} = \frac{108}{12 - 3 - 4 + 1} = 24.02, \quad k - 1 = 2, \quad MSW = 4.33$$

$$F_{2,8} = 5.14, \quad F_{2,8} = 5.14, \quad F_{2,8} = 5.14, \quad F_{2,8} = 5.14$$

Since $F > F_{\alpha}$, we reject H_0 and conclude that there is a difference between the golfer means.

Solutions to"

363

Chapter 20

- 792 - 745

* = -- = 79.2, y = -- = 74.5, 10 J 10

"L x y-(L x)(Ly)

10(63494)-(792)(745) J[lo (77982) - (792) 2][lo (57661) - (745) 2] V(152556)(

44900

21585)

= 0.782

H o :p = 0, H r .p*o r 0.782

1-r 2 l-(0.782) 2 n-2 \ 10-2

= 3.549 , d.f. = n -2 = 10 -2 = 8, t c = 2.306

Since t > t , we reject H a and conclude the correlation coefficient is not equal to zero.

2a & _ »Xr HI»(I» _ 10 (6 3494)-(792)(745) _ 44900

»X* 2 -(X*) 2 ~ 10 (77982)-(792) 2 ~ 152556

« = 3 ^-A* = 74.5-(0.294)79.2= 51.21, ,y = 51.21+0.294*

2b. H o :P = 0, H i: p*o

■■ 0.294

^/'-"^y-b^xy _ /(57661)-51.21(745)-(0.294)(63494)

10-2

10.26

10.26

$$f_{x^2 - nx^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x^2 - \bar{x}^2}{s^2} \right)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{79.2^2 - 79.2^2}{0.294^2} \right)} = 0.0831$$

$$= 0.0831$$

0.0831

$$= 3.538$$

$$r_{f, \cdot} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x^2 - \bar{x}^2}{s^2} \right)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{79.2^2 - 79.2^2}{0.294^2} \right)} = 2.306$$

Since $t > t_c$, we reject H_0 and conclude there is a relationship between payroll and wins.

$$2c. y = 51.21 + 0.294^* = 51.21 + 0.294(70) = 71.79$$

2d

$$CI = y \pm t_{c, s, r}$$

$$[x - x]$$

$$d.f. = n - 2 = 79 - 2 = 77, t = 3.355$$

$$C/ = 71.79 + (3.355)(10.26) = 71.79 + 34.21 = 106.00$$

$$MLO = (79.2)^2$$

$$(77982) -$$

10

$$C/ = 71.79 + (3.355)(10.26)(0.325) = 71.79 + 11.19 = 82.98$$

$$= (0.782)^2 = 0.612 \text{ or } 61.2 \text{ percent}$$

$$J^{**} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x^2 - \bar{x}^2}{s^2} \right)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{79.2^2 - 79.2^2}{0.294^2} \right)} = 3.51$$

Appendi



Statistical Tables

Source: Mr. Carl Schwarz, www.stat.sfu.ca/~cschwarz/. Used with permission.

Binomial Probability Tables

The following table provides the probability of exactly r successes in n trials for various values of p .

Table 1 Binomial Probability Tables

Values of p

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.8100	0.6400	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100
1	0.1800	0.3200	0.4200	0.4800	0.5000	0.4800	0.4200	0.3200	0.1800
2	0.0100	0.0400	0.0900	0.1600	0.2500	0.3600	0.4900	0.6400	0.8100
0	0.7290	0.5120	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010
1	0.2430	0.3840	0.4410	0.4320	0.3750	0.2880	0.1890	0.0960	0.0270
2	0.0270	0.0960	0.1890	0.2880	0.3750	0.4320	0.4410	0.3840	0.2430
3	0.0010	0.0080	0.0270	0.0640	0.1250	0.2160	0.3430	0.5120	0.7290
0	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001
1	0.2916	0.4096	0.4116	0.3456	0.2500	0.1536	0.0756	0.0256	0.0036

2	0.0486	0.1536	0.2646	0.3456	0.3750	0.3456	0.2646	0.1536	0.0486
3	0.0036	0.0256	0.0756	0.1536	0.2500	0.3456	0.4116	0.4096	0.2916
4	0.0001	0.0016	0.0081	0.0256	0.0625	0.1296	0.2401	0.4096	0.6561
0	0.5905	0.3277	0.1681	0.0778	0.0313	0.0102	0.0024	0.0003	0.0000
1	0.3280	0.4096	0.3601	0.2592	0.1563	0.0768	0.0284	0.0064	0.0005
2	0.0729	0.2048	0.3087	0.3456	0.3125	0.2304	0.1323	0.0512	0.0081
3	0.0081	0.0512	0.1323	0.2304	0.3125	0.3456	0.3087	0.2048	0.0729
4	0.0005	0.0064	0.0283	0.0768	0.1563	0.2592	0.3601	0.4096	0.3281
5	0.0000	0.0003	0.0024	0.0102	0.0313	0.0778	0.1681	0.3277	0.5905
0	0.5314	0.2621	0.1176	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000
1	0.3543	0.3932	0.3025	0.1866	0.0938	0.0369	0.0102	0.0015	0.0001
2	0.0984	0.2458	0.3241	0.3110	0.2344	0.1382	0.0595	0.0154	0.0012
3	0.0146	0.0819	0.1852	0.2765	0.3125	0.2765	0.1852	0.0819	0.0146
4	0.0012	0.0154	0.0595	0.1382	0.2344	0.3110	0.3241	0.2458	0.0984
5	0.0001	0.0015	0.0102	0.0369	0.0938	0.1866	0.3025	0.3932	0.3543
6	0.0000	0.0001	0.0007	0.0041	0.0156	0.0467	0.1176	0.2621	0.5314
0	0.4783	0.2097	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	0.0000
1	0.3720	0.3670	0.2471	0.1306	0.0547	0.0172	0.0036	0.0004	0.0000
2	0.1240	0.2753	0.3177	0.2613	0.1641	0.0774	0.0250	0.0043	0.0002

Tables 369

3	0.0230	0.1147	0.2269	0.2903	0.2734	0.1935	0.0972	0.0287	0.0026
---	--------	--------	--------	--------	--------	--------	--------	--------	--------

4	0.0026	0.0287	0.0972	0.1935	0.2734	0.2903	0.2269	0.1147	0.0230
5	0.0002	0.0043	0.0250	0.0774	0.1641	0.2613	0.3177	0.2753	0.1240
6	0.0000	0.0004	0.0036	0.0172	0.0547	0.1306	0.2471	0.3670	0.3720
7	0.0000	0.0000	0.0002	0.0016	0.0078	0.0280	0.0824	0.2097	0.4783
0	0.4305	0.1678	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000
1	0.3826	0.3355	0.1977	0.0896	0.0313	0.0079	0.0012	0.0001	0.0000
2	0.1488	0.2936	0.2965	0.2090	0.1094	0.0413	0.0100	0.0011	0.0000
3	0.0331	0.1468	0.2541	0.2787	0.2188	0.1239	0.0467	0.0092	0.0004
4	0.0046	0.0459	0.1361	0.2322	0.2734	0.2322	0.1361	0.0459	0.0046
5	0.0004	0.0092	0.0467	0.1239	0.2188	0.2787	0.2541	0.1468	0.0331
6	0.0000	0.0011	0.0100	0.0413	0.1094	0.2090	0.2965	0.2936	0.1488
7	0.0000	0.0001	0.0012	0.0079	0.0313	0.0896	0.1977	0.3355	0.3826
8	0.0000	0.0000	0.0001	0.0007	0.0039	0.0168	0.0576	0.1678	0.4305

Poisson Probability Tables

This table provides the probability of exactly x number of occurrences for various values of λ .

Table Z Poisson Probability Tables

Values of λ

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	

0.3679

2 0.0045 0.0164 0.0333 0.0536 0.0758 0.0988 0.1217 0.1438 0.1647
0.1839

3 0.0002 0.0011 0.0033 0.0072 0.0126 0.0198 0.0284 0.0383 0.0494
0.0613

4 0.0000 0.0001 0.0003 0.0007 0.0016 0.0030 0.0050 0.0077 0.0111
0.0153

5 0.0000 0.0000 0.0000 0.0001 0.0002 0.0004 0.0007 0.0012 0.0020
0.0031

1 0.3662 0.3614 0.3543 0.3452 0.3347 0.3230 0.3106 0.2975 0.2842
0.2707

2 0.2014 0.2169 0.2303 0.2417 0.2510 0.2584 0.2640 0.2678 0.2700
0.2707

3 0.0738 0.0867 0.0998 0.1128 0.1255 0.1378 0.1496 0.1607 0.1710
0.1804

4 0.0203 0.0260 0.0324 0.0395 0.0471 0.0551 0.0636 0.0723 0.0812
0.0902

5 0.0045 0.0062 0.0084 0.0111 0.0141 0.0176 0.0216 0.0260 0.0309
0.0361

6 0.0008 0.0012 0.0018 0.0026 0.0035 0.0047 0.0061 0.0078 0.0098
0.0120

7 0.0001 0.0002 0.0003 0.0005 0.0008 0.0011 0.0015 0.0020 0.0027
0.0034

8 0.0000 0.0000 0.0001 0.0001 0.0001 0.0002 0.0003 0.0005 0.0006
0.0009

9 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0001
0.0002

Values of λ_i

x 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0

0 0.1225 0.1108 0.1003 0.0907 0.0821 0.0743 0.0672 0.0608 0.0550
0.0498

1 0.2572 0.2438 0.2306 0.2177 0.2052 0.1931 0.1815 0.1703 0.1596
0.1494

2 0.2700 0.2681 0.2652 0.2613 0.2565 0.2510 0.2450 0.2384 0.2314
0.2240

3 0.1890 0.1966 0.2033 0.2090 0.2138 0.2176 0.2205 0.2225 0.2237
0.2240

4 0.0992 0.1082 0.1169 0.1254 0.1336 0.1414 0.1488 0.1557 0.1622
0.1680

5 0.0417 0.0476 0.0538 0.0602 0.0668 0.0735 0.0804 0.0872 0.0940
0.1008

6 0.0146 0.0174 0.0206 0.0241 0.0278 0.0319 0.0362 0.0407 0.0455
0.0504

7 0.0044 0.0055 0.0068 0.0083 0.0099 0.0118 0.0139 0.0163 0.0188
0.0216

8 0.0011 0.0015 0.0019 0.0025 0.0031 0.0038 0.0047 0.0057 0.0068
0.0081

9 0.0003 0.0004 0.0005 0.0007 0.0009 0.0011 0.0014 0.0018 0.0022
0.0027

10 0.0001 0.0001 0.0001 0.0002 0.0002 0.0003 0.0004 0.0005 0.0006
0.0008

11 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0001 0.0002
0.0002

1 0.1304 0.1135 0.0984 0.0850 0.0733 0.0630 0.0540 0.0462 0.0395
0.0337

2 0.2087 0.1929 0.1771 0.1615 0.1465 0.1323 0.1188 0.1063 0.0948
0.0842

3 0.2226 0.2186 0.2125 0.2046 0.1954 0.1852 0.1743 0.1631 0.1517
0.1404

4 0.1781 0.1858 0.1912 0.1944 0.1954 0.1944 0.1917 0.1875 0.1820 0.1755

5 0.1140 0.1264 0.1377 0.1477 0.1563 0.1633 0.1687 0.1725 0.1747 0.1755

6 0.0608 0.0716 0.0826 0.0936 0.1042 0.1143 0.1237 0.1323 0.1398
0.1462

7 0.0278 0.0348 0.0425 0.0508 0.0595 0.0686 0.0778 0.0869 0.0959
0.1044

8 0.0111 0.0148 0.0191 0.0241 0.0298 0.0360 0.0428 0.0500 0.0575
0.0653

9 0.0040 0.0056 0.0076 0.0102 0.0132 0.0168 0.0209 0.0255 0.0307
0.0363

10 0.0013 0.0019 0.0028 0.0039 0.0053 0.0071 0.0092 0.0118 0.0147
0.0181

11 0.0004 0.0006 0.0009 0.0013 0.0019 0.0027 0.0037 0.0049 0.0064
0.0082

12 0.0001 0.0002 0.0003 0.0004 0.0006 0.0009 0.0013 0.0019 0.0026
0.0034

13 0.0000 0.0000 0.0001 0.0001 0.0002 0.0003 0.0005 0.0007 0.0009
0.0013

14 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0001 0.0002
0.0003 0.0005

15 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001
0.0001 0.0002

Normal Probability Tables

Table 3 provides the area to the left of the corresponding z-score for the standard normal distribution.

Table 3 Normal Probability Tables

Second digit of z

z 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09

0.0 0.5000 0.5040 0.5080 0.5120 0.5160 0.5199 0.5239 0.5279 0.5319
0.5359

0.1 0.5398 0.5438 0.5478 0.5517 0.5557 0.5596 0.5636 0.5675 0.5714
0.5753

0.2 0.5793 0.5832 0.5871 0.5910 0.5948 0.5987 0.6026 0.6064 0.6103
0.6141

0.3 0.6179 0.6217 0.6255 0.6293 0.6331 0.6368 0.6406 0.6443 0.6480
0.6517

0.4 0.6554 0.6591 0.6628 0.6664 0.6700 0.6736 0.6772 0.6808 0.6844
0.6879

2 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09

0.5 0.6915 0.6950 0.6985 0.7019 0.7054 0.7088 0.7123 0.7157 0.7190
0.7224

0.6 0.7257 0.7291 0.7324 0.7357 0.7389 0.7422 0.7454 0.7486 0.7517
0.7549

0.7 0.7580 0.7611 0.7642 0.7673 0.7704 0.7734 0.7764 0.7794 0.7823
0.7852

0.8 0.7881 0.7910 0.7939 0.7967 0.7995 0.8023 0.8051 0.8078 0.8106

0.8133

0.9 0.8159 0.8186 0.8212 0.8238 0.8264 0.8289 0.8315 0.8340 0.8365
0.8389

1.0 0.8413 0.8438 0.8461 0.8485 0.8508 0.8531 0.8554 0.8577 0.8599
0.8621

1.1 0.8643 0.8665 0.8686 0.8708 0.8729 0.8749 0.8770 0.8790 0.8810
0.8830

1.2 0.8849 0.8869 0.8888 0.8907 0.8925 0.8944 0.8962 0.8980 0.8997
0.9015

1.3 0.9032 0.9049 0.9066 0.9082 0.9099 0.9115 0.9131 0.9147 0.9162
0.9177

1.4 0.9192 0.9207 0.9222 0.9236 0.9251 0.9265 0.9279 0.9292 0.9306
0.9319

1.5 0.9332 0.9345 0.9357 0.9370 0.9382 0.9394 0.9406 0.9418 0.9429
0.9441

1.6 0.9452 0.9463 0.9474 0.9484 0.9495 0.9505 0.9515 0.9525 0.9535
0.9545

1.7 0.9554 0.9564 0.9573 0.9582 0.9591 0.9599 0.9608 0.9616 0.9625
0.9633

1.8 0.9641 0.9649 0.9656 0.9664 0.9671 0.9678 0.9686 0.9693 0.9699
0.9706

1.9 0.9713 0.9719 0.9726 0.9732 0.9738 0.9744 0.9750 0.9756 0.9761
0.9767

2.0 0.9772 0.9778 0.9783 0.9788 0.9793 0.9798 0.9803 0.9808 0.9812
0.9817

2.1 0.9821 0.9826 0.9830 0.9834 0.9838 0.9842 0.9846 0.9850 0.9854
0.9857

2.2 0.9861 0.9864 0.9868 0.9871 0.9875 0.9878 0.9881 0.9884 0.9887
0.9890

2.3 0.9893 0.9896 0.9898 0.9901 0.9904 0.9906 0.9909 0.9911 0.9913
0.9916

2.4 0.9918 0.9920 0.9922 0.9925 0.9927 0.9929 0.9931 0.9932 0.9934
0.9936

2.5 0.9938 0.9940 0.9941 0.9943 0.9945 0.9946 0.9948 0.9949 0.9951
0.9952

2.6 0.9953 0.9955 0.9956 0.9957 0.9959 0.9960 0.9961 0.9962 0.9963
0.9964

2.7 0.9965 0.9966 0.9967 0.9968 0.9969 0.9970 0.9971 0.9972 0.9973
0.9974

2.8 0.9974 0.9975 0.9976 0.9977 0.9977 0.9978 0.9979 0.9979 0.9980
0.9981

2.9 0.9981 0.9982 0.9982 0.9983 0.9984 0.9984 0.9985 0.9985 0.9986
0.9986

3.0 0.9987 0.9987 0.9987 0.9988 0.9988 0.9989 0.9989 0.9989 0.9990
0.9990

3.1 0.9990 0.9991 0.9991 0.9991 0.9992 0.9992 0.9992 0.9992 0.9993
0.9993

3.2 0.9993 0.9993 0.9994 0.9994 0.9994 0.9994 0.9994 0.9995 0.9995
0.9995

3.3 0.9995 0.9995 0.9995 0.9996 0.9996 0.9996 0.9996 0.9996 0.9996
0.9997

3.4 0.9997 0.9997 0.9997 0.9997 0.9997 0.9997 0.9997 0.9997 0.9997
0.9998

3.5 0.9998 0.9998 0.9998 0.9998 0.9998 0.9998 0.9998 0.9998 0.9998

0.9998

Tables 373

Student's t-Distribution

Table 4 provides the t-statistic for the corresponding value of alpha or confidence interval and the number of degrees of freedom.

Table *i Student's t-Distribution

Selected right-tail areas with confidence levels underneath

alpha	0.2000	0.1500	0.1000	0.0500	0.0250	0.0100	0.0050	0.0010
	0.0005							

conf

lev	0.6000	0.7000	0.8000	0.9000	0.9500	0.9800	0.9900	0.9980
	0.9990							

d.f.

alpha	0.2000	0.1500	0.1000	0.0500	0.0250	0.0100	0.0050	0.0010
	0.0005							

conf

lev	0.6000	0.7000	0.8000	0.9000	0.9500	0.9800	0.9900	0.9980
	0.9990							

d.f.

Chi-Square Probability Distribution

Table 5 provides the chi-square for the corresponding value of alpha and the number of degrees of freedom.

Table 5 Chi-Square Probability Distribution

Selected right-tail areas

[-Distribution

Table 6 provides the F-statistic for the corresponding degrees of freedom v_1 and v_2 using a value of alpha equal to 0.05.

Table 6 {-Distribution

$\alpha = 0.05$

$\backslash v_1$ 1 2 3 4 5 6 7 8 9 10

1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.882
2	240.543	241.882						

Appendix



Glossary

Addition Rule of Probabilities Determines the probability of the union of two or more events.

Alternative Hypothesis Denoted by H_a represents the opposite of the null hypothesis and holds true if the null hypothesis is found to be false.

Analysis of Variance (ANOVA) A procedure to test the difference between more than two population means.

Bar Chart A data display where the value of the observation is proportional to the height of the bar on the graph.

Bayes' Theorem A theorem used to calculate $P[B/A]$ from information about $P[A/B]$. The term $P[A/B]$ refers to the probability of Event A, given that Event B has occurred.

Biased Sample A sample that does not represent the intended population

and can lead to distorted findings.

Binomial Experiment An experiment that has only two possible outcomes for each trial. The probability of success and failure is constant. Each trial of the experiment is independent of any other trial.

Binomial Probability Distribution A method used to calculate the probability of a specific number of successes for a certain number of trials.

Central Limit Theorem A theorem that states as the sample size, n , gets larger, the sample means tend to follow a normal probability distribution.

Class The interval in a frequency distribution.

Classical Probability Reference to situations when we know the number of possible outcomes of the event of interest.

Cluster Sample A simple random sample of groups, or clusters, of the population. Each member of the chosen clusters would be part of the final sample.

Coefficient of Determination, r^2 Term represents the percentage of the variation in y that is explained by the regression line.

Combinations The number of different ways in which objects can be arranged without regard to order.

Completely Randomized One-Way ANOVA An analysis of variance procedure that involves the independent random selection of observations for each level of one factor.

Conditional Probability The probability of Event A, knowing that Event B has already occurred.

Confidence Interval A range of values used to estimate a population parameter and associated with a specific confidence level.

Confidence Level The probability that the interval estimate will include the population parameter.

Contingency Table A table which shows the actual or relative frequency of two types of data at the same time in a table.

Continuous Random Variable A variable that can assume any numerical value within an interval as a result of measuring the outcome of an experiment.

Correlation Coefficient Indicates the strength and direction of the linear relationship between the independent and dependent variables.

Cumulative Frequency Distribution Indicates the percentage of observations that are less than or equal to the current class.

Data The value assigned to an observation or a measurement and the building block to statistical analysis.

Degrees of Freedom The number of values that are free to be varied given information, such as the sample mean, is known.

Dependent Sample The observation from one sample is related to an observation from another sample.

Glossary J79

Dependent Variable The variable denoted by y in the regression equation that is suspected to be influenced by the independent variable.

Descriptive Statistics Used to summarize or display data so that we can quickly obtain an overview.

Direct Observation Gathering data while the subjects of interest are in their natural environment.

Discrete Probability Distribution A listing of all the possible outcomes of an experiment for a discrete random variable along with the relative frequency or probability.

Discrete Random Variable A variable that is limited to assuming only specific integer values as a result of counting the outcome of an experiment.

Empirical Probability Type of probability that observes the number of occurrences of an event through an experiment and calculates the probability from a relative frequency distribution.

Empirical Rule If a distribution follows a bell-shaped, symmetrical curve centered around the mean, we would expect approximately 68, 95, and 99.7 percent of the values to fall within one, two, and three standard deviations around the mean respectively.

Expected Frequencies The number of observations that would be expected for each category of a frequency distribution, assuming the null hypothesis is true with chi-squared analysis.

Experiment The process of measuring or observing an activity for the purpose of collecting data.

Event One or more outcomes that are of interest for the experiment and which is/ are a subset of the sample space.

Factor Describes the cause of the variation in the data for analysis of variance.

Frequency Distribution A table that shows the number of data observations that fall into specific intervals.

Focus Group An observational technique where the subjects are aware that data is being collected. Businesses use this type of group to gather information in a group setting that is controlled by a moderator.

Fundamental Counting Principle A concept that states if one event can occur in m ways and a second event can occur in n ways, the total number of ways both events can occur together is $m \times n$ ways.

380 Appendix (

Goodness-of-Fit Test Uses a sample to test whether a frequency distribution fits the predicted distribution.

Histogram A bar graph showing the number of observations in each class as the height of each bar.

Hypothesis An assumption about a population parameter.

Independent Event The occurrence of Event B has no effect on the probability of Event A.

Independent Sample The observation from one sample is not related to any observations from another sample.

Independent Variable The variable denoted by x in the regression equation is suspected to influence the dependent variable.

Inferential Statistics Used to make claims or conclusions about a population based on a sample of data from that population.

Interquartile Range Measures the spread of the center half of the data set and is used to identify outliers.

Intersection Two or more events occurring at the same time.

Interval Estimate Provides a range of values that best describe the population.

Interval Level of Measurement Level of data that allows the use of addition and subtraction when comparing values, but the zero point is arbitrary.

Joint Probability The probability of the intersection of two events.

Law of Large Numbers This law states that when an experiment is conducted a large number of times, the empirical probabilities of the process will converge to the classical probabilities.

Least Squares Method A mathematical procedure to identify the linear equation that best fits a set of ordered pairs by finding values for a , the y -intercept; and b , the slope. The goal of the least squares method is to minimize the total squared error between the values of y and \hat{y} .

Level The number of categories within the factor of interest in the analysis of variance procedure.

Level of Significance (α) Probability of making a Type I error.

Glossary 381

Line Chart A display where ordered pair data points are connected together with a line.

Margin of Error Concept determines the width of a confidence interval and is calculated using Z t G s .

Mean Measure is calculated by adding all the values in the data set and then dividing this result by the number of observations.

Mean Square Between (MSB) A measure of variation between the sample means.

Mean Square Within (MSW) A measure of variation within each sample.

Measure of Central Tendency Describes the center point of our data set with a single value.

Measure of Relative Position Describes the percentage of the data below a certain point.

Median The value in the data set for which half the observations are higher and half the observations are lower.

Mode The observation in the data set that occurs most frequently.

Multiplication Rule of Probabilities This rule determines the probability of the intersection of two or more events.

Mutually Exclusive Events When two events cannot occur at the same time during an experiment.

Nominal Level of Measurement Lowest level of data where numbers are used to identify a group or category.

Null Hypothesis Denoted by H_0 , this represents the status quo and involves stating the belief that the mean of the population is $<$, $=$, or $>$ a

specific value.

Observed Frequencies The number of actual observations noted for each category of a frequency distribution with chi-squared analysis.

Observed Level of Significance The smallest level of significance at which the null hypothesis will be rejected, assuming the null hypothesis is true. It is also known as the p-value.

One-Tail Hypothesis Test This test is used when the alternative hypothesis is being stated as $<$ or $>$.

One-Way ANOVA An analysis of variance procedure where only one factor is being considered.

382 Appendix (

Ordinal Level of Measurement This measurement has all the properties of nominal data with the added feature that we can rank the values from highest to lowest.

Outcome A particular result of an experiment.

Outliers Extreme values in a data set that should be discarded before analysis.

α -Value The smallest level of significance at which the null hypothesis will be rejected, assuming the null hypothesis is true.

Parameter Data that describes a characteristic about a population.

Percentiles Measures of the relative position of the data values from dividing the data set into 100 equal segments.

Permutations The number of different ways in which objects can be arranged in order.

Pie Chart Chart used to describe data from relative frequency distributions with a circle divided into portions whose area is equal to the relative frequency distribution.

Point Estimate A single value that best describes the population of interest, the sample mean being the most common.

Poisson Probability Distribution A measurement that is used to calculate the probability that a certain number of events will occur over a specific period of time.

Pooled Estimate of the Standard Deviation A weighted average of two sample variances.

Population A number which represents all possible outcomes or measurements of interest.

Primary Data Data that is collected by the person who eventually uses the data.

Probability The likelihood that a particular event will occur.

Probability Distribution A listing of all the possible outcomes of an experiment along with the relative frequency or probability of each outcome.

Qualitative Data Information which uses descriptive terms to measure or classify something of interest.

Quantitative Data Information which uses numerical values to describe something of interest.

Quartiles Measures the relative position of the data values by dividing the data set into four equal segments.

Glossary 383

Random Variable A variable that takes on a numerical value as a result of an experiment.

Randomized Block ANOVA Analysis of variance procedure that controls for variations from other sources than the factors of interest.

Range Obtained by subtracting the smallest measurement from the

largest measurement of a sample.

Ratio Level of Measurement Level of data that allows the use of all four mathematical operations to compare values and has a true zero point.

Relative Frequency Distribution Displays the percentage of observations of each class relative to the total number of observations.

Sample A subset of a population.

Sample Space All the possible outcomes of an experiment.

Sampling Distribution for the Difference in Means Describes the probability of observing various intervals for the difference between two sample means.

Sampling Distribution of the Mean The pattern of the sample means that will occur as samples are drawn from the population at large.

Sampling Error An error which occurs when the sample measurement is different from the population measurement.

Standard Error of the Difference between Two Means The error describes the variation in the difference between two sample means.

Standard Error of the Estimate (s_e) Measures the amount of dispersion of the observed data around the regression line.

Scheffe Test This test is used to determine which of the sample means are different after rejecting the null hypothesis using analysis of variance.

Secondary Data Data that somebody else has collected and made available for others to use.

Simple Random Sample A sample where every element in the population has a chance at being selected.

Simple Regression A procedure that describes a straight line that best fits a series of ordered pairs (x,y) .

Standard Deviation A measure of variation calculated by taking the square root of the variance.

J84 Appendix (

Standard Error of the Mean The standard deviation of sample means.

Standard Error of the Proportion The standard deviation of the sample proportions.

Statistic Data that describes a characteristic about a sample.

Statistics The science that deals with the collection, tabulation, and systematic classification of quantitative data, especially as a basis for inference and induction.

Stem and Leaf Display This chart displays the frequency distribution by splitting the data values into leaves (the last digit in the value) and stems (the remaining digits in the value).

Stratified Sample A sample that is obtained by dividing the population into mutually exclusive groups, or strata, and randomly sampling from each of these groups.

Subjective Probability This probability is estimated based on experience and intuition.

Sum of Squares Between (SSB) The variation among the samples in analysis of variance.

Sum of Squares Block (SSBL) The variation among the blocks in analysis of vari-

Sum of Squares Within (SSJV) The variation within the samples in analysis of variance.

Surveys Data collection that involves directly asking the subject a series of questions.

Systematic Sample A sample where every k^{th} member of the population is

chosen

for the sample, with value of k being approximately \sqrt{N} , where N equals the size of

the population and n equals the size of the sample. n

Test Statistic A quantity from a sample used to decide whether or not to reject the null hypothesis.

Total Sum of Squares The total variation in analysis of variance that is obtained by adding the sum of squares between (SSB) and the sum of squares within (SSW).

Two-Tail Hypothesis Test This test is used whenever the alternative hypothesis is expressed as \neq .

Type I Error Occurs when the null hypothesis is rejected when, in reality, it is true.

Glossary 385

Type II Error Occurs when the null hypothesis is accepted when, in reality, it is not true.

Union At least one of a number of possible events occur.

Variance A measure of dispersion that describes the relative distance between the data points in the set and the mean of the data set.

Weighted Mean Measure which allows the assignment of more weight to certain values and less weight to others when calculating an average.

B

C

Add-Ins dialog box, 25 addition rule of probabilities,

99-101 alpha role (one sample

hypothesis testing), 231-233 alternative hypothesis, 215 chi-square goodness-of-fit

test, 276 stating, 216-217 analysis of variance. See ANOVA (analysis of variance), 289 Analysis ToolPak, 25 ANOVA (analysis of variance), 289 completely randomized block ANOVA, 301 calculated F-statistic,

303-304 critical F-statistic,

304-305 partitioning the sum of squares, 302-303 one-way ANOVA, 290 completely randomized one-way ANOVA, 291-298 Excel application, 298-299 pairwise comparisons,

299-301 practice, 305-307 assumptions, simple regression, 330 average, 48-50

bar charts, 41-42 bar graphs, 34 Bayes, Thomas, 5

Bayes theorem, 102-103 bell-shape distribution, empirical rule, 69 Bernoulli, James, 5, 121 Bernoulli process. See binomial probability distribution biased samples, 10, 167 BINOMDIST function

(Excel), 127 binomial distribution,

goodness-of-fit test, 280-282 binomial probability distribution (Bernoulli process), 121-126 approximation

normal distribution, 157,

160-161 Poisson distribution, 140-142 characteristics of

experiments, 122-123 Excel calculation, 127-129 mean, 129 practice, 129-130 standard deviation, 129 tables, 126-127 blocking variables, 302

calculated F-statistic

completely randomized

block ANOVA, 303-304 completely randomized one-way ANOVA, 295-296 cells, 283 census data, 17 center angle (pie charts), 40 central limit

theorem, 182-190 central tendency, measures of, 48 Excel application, 56-58 mean, 48-50

mean of grouped data from frequency distribution, 51-54 median, 54-55 mode, 55-56 practice, 58-60 selecting measure, 56 weighted mean, 50-51 characteristics

binomial probability

distribution, 122-123 chi-square distribution,

279-280 normal probability distribution, 146-148 Poisson probability distribution, 132-133 Chart Wizard (Excel), 39 charts, 39

bar charts, 41-42 line charts, 43-44 pie charts, 39-41 Chebyshev's theorem, 71-73

388 The Complete Idiot's Guide to Statistics, Second Edition

chi-square probability distribution, 273 characteristics, 279-280 chi-square statistic, 277 CHINVT function (Excel),

279 critical chi-square score,

277-279 data measurement scales,

274 goodness-of-fit test, 274-275 binomial distribution,

280-282 null and alternative hypothesis, 276 observed versus expected

frequencies, 276-277 practice, 286-288 test for independence, 282-286 class frequencies, 31 classical probability, 82 cluster samples, 171 coefficient of determination,

324-325 combinations (probability),

109-112 complements (probability), 86 completely randomized block ANOVA, 301 calculated F-statistic,

303-304 critical F-statistic, 304-305 partitioning the sum of squares, 302-303 completely randomized one-way ANOVA, 291 calculated F-

statistic,

295-296 critical F-statistic, 296-298 partitioning the sum of squares,
292-295 computer programs

Excel. See Excel (Microsoft)

performance of statistical techniques, 7, 23-26 conditional probability,
94-97 CONFIDENCE function

(Excel), 203-204 confidence intervals, 195 large samples, 196 calculating
intervals,

202-203 changing confidence

levels, 200-201 changing sample size,

201-202 CONFIDENCE function (Excel), 203-204 determining sample
size

for the mean, 202 interpretation, 199-200 interval estimate,

196-198 point estimate, 196 proportion, 208-211 practice, 211-212
regression line, 321-323 small samples, 204-208 confidence levels, 197-
198 construction, frequency

distributions, 31-32 contingency tables, 87, 283 continuous random
variables,

113 CORREL function (Excel),

316 correlation, 311

correlation coefficient, 312-314 calculating with Excel,

315-316 significance, 314-315 practice, 331 counting principles,
probability, 106 combinations, 109-111 Excel applications, 112

fundamental counting principle, 106-107 permutations, 107-109 criteria,
discrete probability

distributions, 115 critical chi-square score,
277-279 critical F-statistic
completely randomized
block ANOVA, 304-305 completely randomized one-way ANOVA, 296-
298 cumulative frequency distribution, 33-34
d.f. (degrees of freedom), 205 data
defined, 15 importance of, 16-17 measurement classification
identification of, 26-27
interval level, 22
nominal level, 21
ordinal level, 21
ratio level, 22-23 ordered pairs. See ordered
pair data presentation, 29
charts, 39-44
frequency distributions, 30-37
practice, 44
stem and leaf display,
37-39
qualitative, 20
quantitative, 20
sources, 17-18

direct observation, 19

experiments, 19

surveys, 20

Index 389

summarization, 47 measures of central

tendency, 48-60 measures of dispersion, 61-78 Data Analysis add-in, 24-26 data measurement scales, chi-square probability distribution, 274 degrees of freedom (d.f.), 205 Deming, W. Edwards (14

points), 6 Department of Commerce,

census data, 17 Department of Labor, labor

statistics, 18 Department of the Interior,

U.S. resource data, 18 dependent events

conditional probability,

96-97 testing difference between means, 263-265 dependent variables, 310-311 descriptive statistics, 6-8 data presentation, 29 charts, 39-44 frequency distributions,

30-37 practice, 44 stem and leaf display, 37-39 data summarization, 47 central tendency, 48-60 dispersion, 61-78 identification of, 12 dialog boxes Add-Ins, 25 ANOVA: Single Factor,

298 Histogram, 36 direct observation, as source

of data, 19 discrete probability distributions, 113

mean, 115-116 rules, 115

standard deviation, 116-118 variance, 116-118 discrete random variables, 113 discrete uniform probability

distribution, 179-180 dispersion, measures of, 61 Chebyshev's theorem,
71-73 Excel calculation, 75-76 measures of relative position, 73-75
practice, 76-78 range, 62-63

standard deviation, 67-71 variance, 63-67 distributions

chi-square probability, 273 characteristics, 279-280 chi-square statistic,
277 CHIINV function

(Excel), 279 critical chi-square score,

277-279 data measurement scales,

274 goodness-of-fit test,

274-282 observed versus expected

frequencies, 276 practice, 286-288 test for independence, 282-286
probability. See probability

distributions sampling, 177-178

central limit theorem,

182-190 mean, 178-180 practice, 193-194 proportion, 190-193 standard
error of the mean, 185-186

E (margin of error), 198 empirical probability, 83 empirical rule

normal probability distribution, 155-156 standard deviation, 69-71 equal
population standard

deviations, 257-260 equations. See formulas error sum of squares (SSE),

293 errors

hypothesis testing, 219-220 sampling, 173 estimators

interval estimate, 196-198 point estimate, 196 ethics, 10-12 events

mutually exclusive, 98-99 probability, 82

intersection of, 87-88 union of, 88-89 Excel (Microsoft) calculations

binomial probabilities,

127-129 central tendency, 56-58 correlation coefficient,

315-316 frequency distributions,

34-37 measures of dispersion,

75-76 normal probabilities,

156-157 permutations and combinations, 112 Poisson probabilities, 139-

140 Chart Wizard, 39 CHINVT function, 279

390 The Complete Idiot's Guide to Statistics, Second Edition

confidence intervals,

203-204 one-way ANOVA, 298-299 performance of statistical
techniques, 7, 23-24 frequency distributions,

34-37 installation of Data Analysis add-in, 24-26 simple regression, 325-
326 TINV function, 241-242 expected frequencies, chi-square probability
distribution, 276-277 experiments

as source of data, 19 binomial characteristics,

122-123 probability, 82

F-distribution (ANOVA), 289 completely randomized

block ANOVA, 301-305 completely randomized one-way ANOVA, 291-
298 Excel application, 298-299 one-way ANOVA, 290 pairwise
comparisons,

299-301 practice, 305-307 factor (ANOVA), 290 focus groups, 19
formulas

classical probability, 82 mean of frequency distribution, 52-53
permutations, 108 population mean, 49 population variance, 65 range,
62 raw score method, 64

sample mean, 48 sample proportion, 191 standard deviation, 67-68
variance, 63

z-score for the proportion, 193 frequency distributions, 30 construction
of, 31-32 cumulative frequency

distribution, 33-34 Excel application, 34-37 histograms, 34 relative
frequency distribution, 32-33 fundamental counting principle,
probability, 106-107

6-ti

goodness-of-fit test chi-square probability distribution, 274-275 binomial
distribution,

280-282 null and alternative hypothesis, 276 Poisson process, 134
Gossett, William, 6, 205 graphs

data presentation

frequency distributions,

30-37 stem and leaf display, 37-39 deceptive presentation, 11-12

Histogram dialog box, 36 histograms, 34 history of statistics, 5-6
hypothesis testing, 213

alternative hypothesis, 215-216

null hypothesis, 215

one sample. See one sample

hypothesis testing one-tail hypothesis test,

218-219, 223-225 practice, 225-226 stating the null and alternative
hypothesis, 216-217 two samples. See two sample hypothesis testing two-

tail hypothesis test, 217-218 sample, 220-223 scale of the original variable, 221-222 standardized normal scale, 222-223 Type I and II errors, 219-220

l-J-K

independence test, chi-square

distribution, 282-286 independent events, conditional probability, 96-97 independent samples, 265-269 independent trials, 5 independent variables,

310-311 inference, 5

inferential statistics, 6-9 correlation, 311-316 hypothesis testing, 213 alternative hypothesis,

215-216 null hypothesis, 215 one sample, 227-247 one-tail hypothesis test,

218-219, 223-225 practice, 225-226 stating the null and alternative hypothesis, 216-217 two samples, 250-270

Index 391

two-tail hypothesis test,

217-223 Type I and II errors, 219-220 identification of, 12 independent versus dependent variables, 310-311 samples, 165 cluster, 171 errors, 173 poor techniques, 174 population versus,

166-167 practice, 176 random, 167-170 sampling distributions,

177-194 stratified, 172 systematic, 170 simple regression, 316 assumptions, 330 coefficient of determination, 324-325 Excel application,

325-326 least squares method,

317-321 multiple regression

versus, 330-331 negative correlation example, 326-330 regression line confidence intervals, 321-323 regression line slope, 323-324 information, data versus, 17 Internet, misuse of statistics,

12 interquartile range (IQR),

74-75 intersection of events,

probability, 87-88 interval estimates, 196-198 interval level of measurement (data), 22

intervals, frequency distributions, 31 IQR (interquartile range),

74-75

L

labor statistics, 18

least squares method (linear

equations), 317-321 level (ANOVA), 291 line charts, 43-44 linear relationships

correlation coefficient,

312-316 simple regression, 316 assumptions, 330 coefficient of determination, 324-325 Excel application,

325-326 least squares method,

317-321 multiple regression

versus, 330-331 negative correlation example, 326-330 regression line confidence intervals, 321-323 regression line slope, 323-324

M

margin of error (£), 198 mean, 48-50

binomial probability

distribution, 129 confidence intervals for large samples, 196 calculating intervals,

202-203 changing confidence levels, 200-201

changing sample size,

201-202 CONFIDENCE function (Excel), 203-204 determining sample size,

202 interpretation, 199-200 interval estimate,

196-198 point estimate, 196 confidence intervals for small samples, 204-208 discrete probability distributions, 115-116 sampling distributions,

178-180 two sample hypothesis testing, 252-255 dependent samples,

263-265 equal population standard deviations, 257-260 small sample size and unknown sigma, 256 unequal population standard deviations, 260-263 mean square between (MSB),

296 mean square within (MSW),

296 measurements

central tendency, 48

Excel application, 56-58 mean, 48-50 mean of grouped data from frequency distribution, 51-54 median, 54-55 mode, 55-56 practice, 58-60 selecting measure, 56 weighted mean, 50-51 data, 21

392 The Complete Idiot's Guide to Statistics, Second Edition

identification of, 26-27 interval level, 22 nominal level, 21 ordinal level, 21 ratio level, 22-23 dispersion, 61

Chebyshev's theorem,

71-73 Excel calculation, 75-76 measures of relative

position, 73-75 practice, 76-78 range, 62-63 standard deviation,

67-71 variance, 63-67 relative position, 73-75 median, 54-55 Microsoft Excel. See Excel

(Microsoft) mode, 55-56 MSB (mean square between),

296 MSW (mean square within),

296 multiple regression, 330-331 multiplication rule of probabilities, 97-98 mutually exclusive classes (frequency distributions), 32 mutually exclusive events, 98-99

negative correlation, 326-330 negative linear correlation,

311 Nielsen Media Research, 166 nominal level of measurement

(data), 21 normal probability distribution, 145 approximating binomial distribution, 157, 160-161

calculating probabilities, 148 empirical rule, 155-156 Excel, 156-157 standard normal table,

150-155

standard z-score,

148-150

characteristics, 146-148

practice, 161-162

NORMDIST function

(Excel), 156 null hypothesis, 215

chi-square goodness-of-fit

test, 276 stating, 216-217 numerical summarization of data, 47 central tendency, 48

Excel application, 56-58 mean, 48-50 mean of grouped data from

frequency distribution, 51-54 median, 54-55 mode, 55-56 practice, 58-60 selecting measure, 56 weighted mean, 50-51 dispersion, 61

Chebyshev's theorem,

71-73 Excel calculation, 75-76 measures of relative

position, 73-75 practice, 76-78 range, 62-63 standard deviation,

67-71 variance, 63-67

object order, permutations, 107-109

observation, direct, 19 observed frequencies, 276-277 observed level of significance,

233 one sample hypothesis testing, 227 large sample when sigma is

known, 228-229 large sample when sigma is

unknown, 229-231 one-tail hypothesis test for

proportion, 243-245 p-value, 233-236 practice, 246-247 proportion with large samples, 242 role of alpha, 231-233 small sample when sigma is

known, 236-237 small sample when sigma is

unknown, 237-241 TINV function (Excel),

241-242 two-tail hypothesis test for proportion, 245-246 one-tail hypothesis test, 218-219 p-value, 233-234 sample, 223-225 one-way ANOVA, 290 completely randomized one-way ANOVA, 291 calculated F-statistic,

295-296 critical F-statistic,

296-298 partitioning the sum of squares, 292-295 Excel application, 298-299 ordered pair data, 310 correlation, 311-316 independent versus dependent variables, 310-311 simple regression, 316 assumptions, 330

Index 393

coefficient of determination, 324-325 Excel application,
325-326 least squares method,
317-321 multiple regression
versus, 330-331 negative correlation example, 326-330 regression line
confidence intervals, 321-323 regression line slope, 323-324 order of
objects, permutations,
107-109 ordinal level of measurement (data), 21
p-value, 233-236
pairwise comparisons, Scheffe
test, 299-301 parameters (population), 16,
197 partitioning the sum of squares completely randomized
block ANOVA, 302-303 completely randomized one-way ANOVA, 292-
295 permutations, probability,
107-109, 112 Petty, Sir William, 5 pie charts, 39-41 point estimate, 196
Poisson, Simeon, 131 POISSON function (Excel),
139 Poisson probability distribution, 131
approximating binomial
distribution, 140-142 calculating with Excel,
139-140 characteristics, 132-133 practice, 142-143 tables, 136-139
population mean, 49 parameters, 16 sample versus, 8, 166-167 variance,
65-67 positive linear correlation, 311 posterior probabilities, 96
presentation of data, 29 charts, 39
bar charts, 41-42 line charts, 43-44 pie charts, 39-41 graphs
frequency distributions,

30-37 stem and leaf display, 37-39 practice, 44 primary sources of data, 17-20 prior probabilities, 95 probability, 101

addition rule, 99-101 Bayes theorem, 102-103 classical, 82 conditional, 94-97 counting principles, 106 combinations, 109-111 Excel applications, 112 fundamental counting

principle, 106-107 permutations, 107-109 defined, 82 distributions, 112

binomial experiments,

121-130 discrete, 113-118 normal distribution, 145-162

Poisson process, 131-143 practice, 118-119 random variables, 112-113 empirical, 83 intersection of events,

87-88 multiplication rule, 97-98 mutually exclusive events,

98-99 posterior probabilities, 96 practice, 89-91, 103-104 prior probabilities, 95 properties, 86-87 subjective, 85 union of events, 88-89 proportion

confidence intervals for

large samples, 208-211 one sample hypothesis

testing, 242 one-tail hypothesis test,

243-245 sampling distributions,

190-193 two sample hypothesis

testing, 265-269 two-tail hypothesis test,

245-246 purpose of statistics, 4

M.

qualitative data, 20 quality control, 6 quantitative data, 20 quartiles, 73-74

r (correlation coefficient), 313 RAND function (Excel), 170 random number table, 169 random samples, 167-170 random variables

Poisson probability distribution, 132

39*t The Complete Idiot's Guide to Statistics, Second Edition

probability distributions, 112-113 range (measure of dispersion),

62-63 ratio level of measurement

(data), 22-23 raw score method (variance

calculations), 64-65 relative position, 73-75 regression lines

confidence intervals,

321-323 slope, 323-324 simple regression. See simple regression relative frequency distribution,

32-33 resources, data, 18 role of alpha, one sample hypothesis testing, 231-233

samples biased, 10 clustered, 171 confidence intervals for the mean, 196 calculating intervals,

202-203 changing confidence

levels, 200-201 changing sample size,

201-202 CONFIDENCE function (Excel), 203-204 determining sample size,

202 interpretation, 199-200 interval estimate,

196-198 point estimate, 196 practice, 211-212 proportion, 208-211 small samples, 204-208

distributions, 177

central limit theorem,

182-190 mean, 178-180 practice, 193-194 proportion, 190-193 standard error of the mean, 185-186 errors, 173 mean, 48

poor techniques, 174 population versus, 8,

166-167 practice, 176 random, 167-170 stratified, 172 systematic, 170 SAS, 7 scale of the original variable,

221-222 Scheffe test, 299-301 secondary sources of data,

17-18 significance, correlation coefficient, 314-315 simple random sampling,

168-170 simple regression, 316 assumptions, 330 coefficient of determination, 324-325 Excel application, 325-326 least squares method,

317-321 multiple regression versus,

330-331 negative correlation example, 326-330 practice, 331 regression line, 321-324 slope, regression line, 323-324 software

Excel. See Excel (Microsoft) performance of statistical techniques, 7, 23-26

sources of data, 17-20

SPSS, 7

SSB (sum of squares between),

293 SSBL (sum of squares block),

303 SSE (error sum of squares),

293 SST (total sum of squares),

294 SSW (sum of squares within),

303 standard deviation, 67

binomial probability distribution, 129 discrete probability distributions,

116-118 empirical rule, 69-71 grouped data calculation, 67-69 standard error of the mean,

185-186 standard error of the proportion, 192-193 standard normal distribution,

149 standard normal table (normal probability distribution),

150-155 standard z-score (normal probability distribution),

148-150 standardized normal scale,

222-223 Statistics Canada, 18 stem and leaf display, 37-39 stratified samples, 172 subjective probability, 85 summarization of data. See

numerical summarization of

data surveys, as source of data, 20 symmetrical curve distribution, 69 systematic samples, 170

Index 395

t-distribution, 205

t-test, 6

tables

binomial probability,

126-127 contingency, 87 Poisson probability distribution, 136-139 random number, 169 standard normal distribution, 150-155 test for independence,

282-286 theoretical sampling distribution of the mean, 186-188 TINV function (Excel),

241-242 total sum of squares (SST),

294 trials, independent, 5 true zero point, 22 twice as much rule, 22 two sample hypothesis testing, 250 differences between means, 252 dependent samples,

263-265 equal population standard deviations, 257-260 small sample size and unknown sigma, 256 unequal population standard deviations, 260-263 differences between proportions, 265-269 differences other than zero,

255-256 practice, 269-270

sampling distribution for the difference in means, 250-252 two-tail hypothesis test, 217-218 p-value, 234-236 sample, 220-223 scale of the original variable, 221-222 standardized normal scale, 222-223 Type I errors, 219-220 Type II errors, 219-220

u-v-w

U.S. resource data, 18 unequal population standard

deviations, 260-263 union of events, probability,

88-89

variables, 310-311 variance (measure of dispersion), 63 discrete probability distributions, 116-118 population variance, 65-67 raw score method, 64-65

weighted mean, 50-51

X-Y-Z

x-axis (line charts), 44 y-axis (line charts), 44

Dear Reader,

Welcome to my world of statistics! I want to commend you for seeking help with this very challenging topic. Countless individuals out there like

you are struggling with statistics, and many of those don't make the effort to seek additional help. I, too, was nearly one of those statistics (sorry, I just love to use that word!) back in my graduate school days. One of my required courses was an advanced, theoretical statistics class with seven students that was taught by a very nice professor who was a brilliant researcher with only one minor flaw—the man couldn't teach you how to lick a stamp. After two classes, a feeling of panic started to set in as I saw my dreams of earning a Ph.D. fading away. My predominant thought in class was, "What is this guy talking about?"

Like you, I decided to seek help. Unfortunately for me, the Complete Idiot's Guide series hadn't been invented yet. So I sought the help of a private tutor. Eugene was an international graduate student with a limited ability to speak English, but he had a phenomenal sense of explaining abstract concepts. I quickly fell into the routine of leaving class in a complete fog, meeting with Eugene, and then exclaiming, "Eureka, that's what he was talking about!" I went on to receive an "A" in this class, earned my degree, and the rest is history. As a token of my appreciation to Eugene, I presented him with my first-born male child. (I'm only kidding, Brian!)

Based on my own experiences, my advice to you is to either find a brilliant international graduate student with a limited ability to speak English who can explain abstract concepts with amazing clarity to personally tutor you, or use this book.

Each statistical concept in the chapters that follow is explained in loving detail with plenty of examples and, when appropriate, a little humor. In writing this book, my goal has been to play the role of Eugene for you and explain those messy concepts in a way that makes sense to you, so you can say "Eureka!" Only I won't cost as much as Eugene, and I hope my English is a little better.

Bob Donnelly /

About the Author

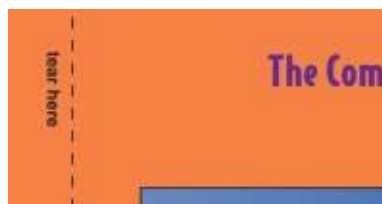
Robert A. Donnelly, Jr., Ph.D. (bob@stat-guide.com) is a professor at Goldey-Beacom College in Wilmington, Delaware, with more than 20

years of teaching experience. He teaches classes in statistics, operations management, management information systems, and database management at both the undergraduate and graduate level. Bob earned an undergraduate degree in chemical engineering from the University of Delaware, after which he worked for several years as an engineer in a local chemical plant. Despite success in this field, Bob felt drawn to pursue a career in education. It was his desire to teach (or maybe he just had a bad day) that took him back to school to earn his MBA and Ph.D. in operations research, also from the University of Delaware. Go Blue Hens!

Bob's working experience prior to his teaching career provides him with many opportunities to incorporate real-life examples into classroom learning. His students appreciate his knowledge of the business world as well as his mastery of the course subject matter. Many former students seek Bob's assistance in work-related issues that deal with his expertise. Typical student comments focus on his genuine concern for their welfare and his desire to help them succeed in reaching their goals. They also love when he cancels class because the roads in his backwoods neighborhood have flooded.

While keeping teaching as his main focus, Bob performs consulting activities through his firm, Partners for Strategic Solutions, which provides services for businesses seeking management techniques to improve performance. He recently completed a test bank for a new textbook on mathematical modeling using Excel for Prentice-Hall Publishers. Bob has also remained current with today's technology with CIW certification as Master CIW Designer. You can reach him at bob@stat-guide.com.

It is obvious to anyone that Bob's first love is teaching. His children can attest to that when his eyes light up at the end of the day and he asks "Well, does anybody need help with their math homework?" Sometimes they say yes just to make him happy.



The Complete Idiot's Guide® to Statistics Reference Card